UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

EXTREME VALUE ANALYSIS OF COMPETITIVE FREEDIVING RECORDS

Mestrado em Estatística e Investigação Operacional

Especialização em Estatística

Jessica Silva Lomba

Dissertação orientada por:

Professora Doutora Maria Isabel Fraga Alves

2016

Acknowledgments

First and foremost, to my parents, that gave me life and everyday invest deeply on my formation, both personal and academic. I would not be where I am today without their unwavering support, affection and belief in me.

To my advisor, Professor Maria Isabel Fraga Alves, who I have come to know and admire during the last two years. Her guidance, patience, support, availability and commitment to helping me were key pieces in the accomplishment of this work, as was all the knowledge on the subject that Professor Isabel bestowed upon me. It was truly a pleasure and a privilege to work with such a delightful and insightful person, whose friendship I have come to treasure.

To Pedro, for always being there, no matter the distance, when it was hardest just as when it was easiest. For dealing with my lows with the same profound love as with my ups. For giving me something to hope for, keeping me focused and believing in me more than I did myself.

To all my friends, who never allowed me to give up and helped me keep moving forward. A special word goes to some particular people who played a direct part on my endeavour through this Masters course: to Raquel and Daniel, for all the patience and help they graced me with, leading the way and sharing what they had learned, from academic discoveries, through mastering LaTeX and down to the latest piece of gossip; to Diogo and Tiago, my "little boys" who were always there even when they weren't, proving that family is not defined by blood; to the girls at House Yona, Bruna, Ana, Filipa and Inês, whose hospitality was more than I could ever ask for, making sure I always felt at home in my work trips to Lisbon; to Joana, for the true companionship we shared through this phase; to Patrícia and Isa, for the good spirits and inspirational incentive they always send my way, regardless of how far apart we are. And to everyone I did not name but who know they hold a very important space in my heart and life.

To all of you, my most sincere thank you. You all are the reason why this was worth it.

Obrigada.

Do not breathe under the water.

Abstract

In this dissertation we introduce the basic concepts of Extreme Value Theory (EVT), presenting the theoretical principals it is built on, detailing the most common statistical methodologies used for this type of data, and finally illustrating the approaches by the analysis of a concrete data set.

Extreme Value Theory is the field of Statistics that deals with extremal occurrences, that is, with very large or very small events that occur rarely and for which the samples are therefore scarce. Unlike Classical Statistics, which is focused in modelling the bulk of central data, EVT focuses in understanding the behaviour of observations that fall furthest from the centre of the sample, allowing for the extrapolation of conclusions beyond the previously observed data.

EVT is based in the most fundamental theorems by Fisher and Tippett (1928), Gnedenko (1943), Pickands (1975) and Balkema and de Haan (1974), with great influence of the distributional theory of Order Statistics and Regular Variation theory. Some assumptions are usually made for the application of such results, the most common being that of independence and identical distribution of the random variables in the sample. However, this assumptions can sometimes be relaxed, as it is the case when considering non-stationary data. The main inference methodologies considered are the parametric approach, here comprised by the Gumbel method, the Peaks Over Threshold method and the Largest Observations method, and the semi-parametric approach. Literary references will be provided throughout the text for further information on the addressed topics and related application examples.

This framework has been proven useful in various fields, with particular visibility in Hydrology, Environmental Sciences, Finances and Engineering. It has also been used in Sports data, as is the case in this work.

Freediving is an international competitive sport that revolves around the divers' capability of holding their breaths (apnea) underwater without the aid of oxygen tanks or breathing tubes. The Static Apnea modality consists in recording the maximum time the freediver holds their breath with his nose and mouth immersed while floating on the surface of the water or standing on the bottom of a pool.

The data analysed in the Case Study of this dissertation consists of the best personal records of female freedivers in the Static Apena modality that achieved at least a 3 minute breath hold. The sample maximum corresponds to the current female world record of 9 minutes and 2 seconds, and there is an interest in estimating the probability that this mark will be overcome. Another

interesting factor to determine is the existence of a maximum limit statistically possible for the apnea time of a member of this population – that would speak to the existence of such a limit for the general female population.

Both parametric and semi-parametric approaches will be used on the data with the aim of estimating extremal quantiles and exceedance probabilities, and evaluating the finiteness of the right endpoint. Finally, the veracity of the stationarity assumption from which the analysis starts will also be tested.

Keywords: Extreme Value Theory, Parametric Inference, Semi-parametric Inference, Non-Stationarity, Freediving.

Resumo

Nesta dissertação serão introduzidos os conceitos básicos inerentes à Teoria de Valores Extremos, começando pela apresentação dos princípios teóricos sobre os quais é construída, passando pela exposição das metodologias estatísticas mais comuns para tratar este tipo de dados, e finalizando com a exemplificação das abordagens mencionadas com a análise de um conjunto de dados reais.

A Teoria de Valores Extremos é o campo da Estatística especializado em lidar com ocorrências extremas dos processos, ou seja, com os valores muito elevados ou muito reduzidos que raramente se registam, razão pela qual as amostras são geralmente escarças. Ao contrário do que acontece com a Teoria Estatística Clássica, que se foca na modelação do grande conjunto de dados centrais, a Teoria de Valores Estremos foca-se na compreensão do comportamento das observações que se registam o mais afastado do centro da amostra, permitindo a extrapolação das conclusões obtidas para além dos dados anteriormente observados (estimação além da amostra).

A Teoria de Valores Extremos baseia-se fundamentalmente nos teoremas trabalhados por Fisher and Tippett (1928), Gnedenko (1943), Pickands (1975) e Balkema and de Haan (1974), sendo uma área com grandes influências da teoria distribucional (exata e assintótica) das Estatísticas Ordinais e ainda da teoria da Variação Regular. A utilização dos resultados apresentados depende geralmente de algumas suposições necessárias, sendo que a mais comum é a de independência e de idêntica distribuição das variáveis aleatórias que compõem a amostra. No entanto, é possível relaxar estas suposições de variadas formas, como por exemplo é o caso do tratamento de dados não estacionários (cuja variação do tempo, ou outras variáveis exógenas, influenciam a distribuição).

Do ponto de vista estacionário, serão abordadas as principais metodologias que constituem uma abordagem paramétrica, assim como uma abordagem geral semi-paramétrica. Será também referida, embora de forma mais breve, uma metodologia possível para lidar com a não estacionariedade temporal.

Este tipo de análise de valores extremos já se provou útil e até indispensável em várias áreas de conhecimento, com especial visibilidade nos campos da Hidrologia (tratamento de dados sobre níveis máximos da água do mar ou níveis de precipitação que podem causar cheias desastrosas), Ciências Ambientais (estudo das temperaturas globais ou das velocidades do vento que mostram a alteração das condições climáticas extremas no planeta), Finanças (estudo de períodos de retorno e probabilidades de *crash*) ou Engenharia (inferência sobre resistência de materiais ou fiabilidade de equipamentos). Outra área de interesse em que se aplicam estas metodologias é

a área do Desporto, em que, dependendo da modalidade, se lida com distâncias, tempos, pesos, etc., máximos ou mínimos, mas em que são invariavelmente os extremos, os recordes, que se destacam. Este é o âmbito em que se inserem os dados tratados no Case Study deste trabalho.

O Mergulho em Apneia, ou *Freediving*, é um desporto de competição internacional que testa a capacidade dos mergulhadores em suster a respiração (i.e. permanecer em apneia) debaixo de água, sem recurso a tanques de oxigénio ou tubos de respiração. É também proibida a preparação dos atletas com a respiração de oxigénio puro antes das provas.

Existem 8 modalidades reconhecidas pela entidade reguladora deste desporto, a AIDA – *Association Internationale pour le Développement de l'Apnée*. À exceção da modalidade de Apneia Estática, todas as restantes consistem em medir a distância máxima percorrida pelo mergulhador em apneia, sob diferentes condições (com ou sem babatanas, com ou sem cabo, entre outras). A modalidade de Apneia Estática consiste em cronometrar o tempo máximo que o mergulhador sustém a respiração com as vias respiratórias (nariz e boca) submersos, enquanto flutua à superfície da água ou de se encontra em pé no fundo da piscina. As competições, normalmente realizadas em piscinas interiores artificiais, podem também ser realizadas no mar, em águas rasas e calmas.

O conjunto de dados que será o objeto da análise do Case Study desta dissertação insere-se nesta modalidade, e consiste no melhor registo pessoal de mergulhadoras de competição femininas que conseguiram marcas de no mínimo 3 minutos em Apneia Estática. Os dados, referentes ao período entre os anos 2002 e 2014, estão disponíveis ao público *online* no sítio oficial da AIDA, mas não é conhecida nenhuma outra análise estatística do género feita sobre estes registos.

O máximo da amostra sob análise corresponde ao recorde feminino atual de 9 minutos e 2 segundos de submersão estática, e foi conseguido pela várias vezes campeã mundial Natalia Molchanova, no ano de 2013. Esta nadadora faleceu tragicamente em 2015, no exercício do desporto, enquanto dava uma aula de mergulho em apneia em mar aberto. Este facto motivou a escolha destes dados para a ilustração das técnicas de Teoria de Valores Extremos.

Será interessante avaliar vários indicadores relativos a estes dados, como por exemplo estimar a probabilidade de que o atual recorde mundial feminino venha a ser ultrapassado. Probabilidades muito reduzidas indicam que foi atingida uma zona de estabilidade em que dificilmente poderão ser obtidas melhores marcas.

Outra característica inerente à população com grande importância a determinar será a existência (ou não) de um limite máximo estatisticamente possível para o tempo de apneia de um membro desta população de competição. A existência de um tal limite poderá levar a conclusões imediatas acerca do tempo máximo durante o qual um membro da população feminina em geral poderá, no limite, suster a respiração.

Sob a condição de estacionariedade, para a obtenção de estimativas para estes e outros indicadores, serão aplicadas três das mais comuns abordagens em contexto paramétrico: a abordagem de Gumbel, também conhecida como método dos Máximos por Bloco, apoiada no Teorema de Fisher-Tippet-Gnedenko e que faz uso da distribuição Generalizada de Valores Extremos para modelar o conjunto dos máximos de cada subamostra ou bloco; a abordagem Paretiana de Exces-

sos, em inglês *Peaks Over Threshold*, que apoiada pelo Teorema de Pickands-Balkema-de Haan modela os excessos acima de um determinado nível com uma distribuição Generalizada Pareto; a abordagem das Maiores Observações, que com base na distribuição assintótica conjunta das k observações de topo visa ajustar um modelo Generalizado de Valores Extremos Multivariado à amostra das k maiores observações de cada subamostra ou bloco.

Estes métodos paramétricos são passíveis de adaptação para casos específicos de não estacionariedade. Tal será ilustrado também sobre a amostra de recordes de Mergulho em Apneia Estática, em especial para o caso em que se procura inferir sobre a existência de tendência ao longo do tempo nos dados. Outro tipo de não estacionariedade, que não será aqui abordado, consiste na existência sazonalidade nos dados, recorrente em variáveis ambientais, por exemplo, onde as estações influenciam de forma diferente as séries de dados.

A abordagem semi-paramétrica será também exemplificada sobre os dados do Case Study, apenas sob a suposição de estacionariedade dos dados. Esta não se baseia em qualquer modelo ou distribuição, mas depende unicamente do tipo de comportamento da cauda da distribuição subjacente e desconhecida que os dados sugerem. Serão apresentados alguns dos estimadores mais usuais neste contexto semi-paramétrico e introduzidas as condições teóricas mais gerais que devem ser verificadas para a sua conveniente aplicação.

Os métodos abordados ao longo desta dissertação não são de qualquer modo extensivos ou restritivos de toda a Análise de Valores Extremos. Trata-se de um vasto campo ainda em desenvolvimento e com grande interesse atualmente, sendo possível encontrar na literatura uma miríade de informação relativa quer a fundamentos teóricos quer a novas metodologias estatísticas que expandem os limites desta área. Ao longo do texto serão sempre citados autores e trabalhos para os quais se remetem para mais informação e fundamentação acerca dos tópicos abordados.

Nesta dissertação pretende-se apenas levantar a ponta do véu da Teoria de Valores Extremos, demonstrando de forma breve a sua utilidade com o objetivo de incentivar a curiosidade sobre o tema. Pretende-se ainda, simultaneamente, fazer uma análise inicial a um conjunto de dados interessante sobre um tema relacionável e que se encontrava "a cru" até aqui.

Palavras-Chave: Teoria de Valores Extremos, Inferência Paramétrica, Inferência Semi-Paramétrica, Não Estacionaridade, Mergulho em Apneia.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

ACF: *Autocorrelation Function*
AIDA: *Association Internationale pour le Développement de l'Apnée*
BM: *Block Maxima*
CI: *Confidence Interval*
CLT: *Central Limit Theorem*
d.f.: *Distribution Function*
EVA: *Extreme Value Analysis*
EVI: *Extreme Value Index*
EVT: *Extreme Value Theory*
GEV: *Generalized Extreme Value Distribution*
GP: *Generalized Pareto Distribution*
i.i.d.: *Independent and Identically Distributed*
LAN: *Locally Asymptotically Normal*
LO: *Largest Observations*
LRT: *Likelihood Ratio Test*
m.e.f.: *Mean Excess Function*
ML: *Maximum Likelihood*
NSBM: *Non-Stationary Block Maxima*
NSLO: *Non-Stationary Largest Observations*
NSPOT: *Non-Stationary Peaks Over Threshold*
o.s.: *Order Statistics*
p.d.f.: *Probability Density Function*
POT: *Peaks Over Threshold*
PORT: *Peaks Over Random Threshold*
PWM: *Probability Weighted Moments*
r.v.: *Random Variable*
SA: *Static Apnea*
se: *Standard Error*

Chapter 1

Introduction

In Classical Statistical Theory, we are presented with a collection of events and seek a possible family of distributions and its specific parameters to fit the variable that produced that sample. To this end, we focus our attention on the bulk of central data, searching for a distribution that suits the majority of observations that concentrate around the sample median. In this approach, extremal values, i.e., observations that fall furthest from the centre of the sample, can often be considered outliers and so discarded from the analysis.

This is an acceptable procedure if we aim to infer about the most common events, the everyday observations. But we know better than to believe that every real world process can be interpreted in such a simplistic way.

Especially in natural processes, such as the sea or precipitation levels (examples from the field of Hydrology), it is often that the extreme and rare events are the ones that cause great damages, like floods or droughts. In the Financial world, extremal observations of asset prices or interest rates can have a major impact on Economy, and we can think as well about the effect that an abnormally big insurance prize would have on the accounting of an unprepared insurance company. Even in Engineering, the resistance of the isolated materials can determine the life span of the final product they compose. These examples seldom show themselves on the samples of the respective random variables (r.v.), for they are extreme and rare.

There is an example through which we can easily comprehend the necessity for a different probabilistic and statistical theory, one which better deals with these extremal occurrences than the classical one. It is the same example that lead the pioneer of the approach to this same conclusion. In the 1920's Leonard Tippett was asked by his employer, the British Cotton Industry Research Association, to find a way to improve the strength of cotton thread. Realizing the thread was only as strong as its weakest fibres, and with the assistance of Sir Ronald Fisher, he developed an asymptotic theory that restricted the behaviour of the distribution function in the tail, that is, of the very large (or very small) values, to belong to one of three classes of distribution functions that can be fitted to that extremal part of the sample – Fisher and Tippett (1928). And thus it had been set the foundation for the field of Statistics now known as Extreme Value Theory (EVT), where the probabilistic structure of order statistics (o.s.) has high importance, as will be shown.

Other early significant contributions for the development of EVT came from Fréchet (1927), von Mises (1936) and Gnedenko (1943). Relevant influences to the statistical approach to this theory came from Gumbel (1985), who made an important advance in promoting the use of EVT as a vital tool in the analysis of extremal behaviour of physical processes, with the publishing of his book *Statistics of Extremes*; Pickands (1975) and Balkema and de Haan (1974) also represent unavoidable references in EVT. We should note that this subject continues to be the focus of the work not only of many statisticians nowadays, but also of researches and practitioners that are faced with extreme value problems in their fields.

Since the aim of EVT is to describe the abnormal rather than the normal events, the focus is set on the tails of the distribution. Frequently, in this framework, there is high scarcity of the events we wish to infer about, and it is often that there aren't records of such rare happenings. But of course, just because we haven't seen it happen, it should not mean these events are impossible, as they in many cases certainly aren't. Take, for example, how rare it is for a significantly strong earthquake to be felt in Continental Portugal. If we had a sample of 100 year records of seismic activity felt in the country, say dating from 1600 to 1700, we could be lead to believe that the earthquake that nearly destroyed Lisbon on November 1st 1755 would be impossible, just because similar records did not exist. The EVT provides us here with a strong theoretical and statistical basis for extrapolating beyond the sample.

The same principal can be applied to sports records. Every time a new best time, or best distance, or best weight record is set, there was by definition no prior knowledge of an observation that extreme. Regardless, new bests are accomplished regularly in the world of competitive sports, and sometimes the pursuit of this improbable results can culminate in serious consequences for the athletes.

On 2nd August 2015, Freediving world champion Natalia Molchanova tragically died on the Mediterranean Sea while giving a lesson of the sport. Her death was not the first related to the modality, and true as it may be that she did not die trying to break a record, her accident brought more awareness to the dangers that this sport entails, which can include forcing the body to accomplish unreasonable submersion times.

Freediving is an international competitive sport that revolves around the divers' capability of holding their breaths (apnea) underwater without the aid of oxygen tanks or breathing tubes. There are 8 official competitive disciplines that qualify under the freediving sport, regulated by AIDA (Association Internationale pour le Développement de l'Apnée): no limit; variable weight; constant weight; constant weight without fins; free immersion; dynamic with fins; dynamic without fins; static apnea (SA). This last one, the focus of the case study ahead, consists in "the freediver holding his breath for as long as possible with his nose and mouth immersed while floating on the surface of the water or standing on the bottom of a pool", as defined in AIDA (2016), and it is the only one of the 8 disciplines that regards time instead of distance. The point is to achieve the maximum time possible, and therefore extreme events play a key part in analysing data registered from this competitions.

AIDA has been recording performances of divers that participate in their competitions since the late 1990's, in the several disciplines of the sport, separated by gender. The current SA

records belong to the Frenchman Stéphane Mifsud, who was submersed for 11 minutes and 35 seconds in 2009, and precisely to the Russian Natalia Molchanova that held her breath for 9 minutes and 2 seconds in 2013. Are this records unbreakable, given the current state of the art? Is there a limit to the maximum apnea time possible, statistically speaking? These are some of the questions we are concerned with and which we will use the EVT techniques to try to answer. It is important to note that biological and physiological arguments will not be used, and that the approach to the problem will be purely statistical.

Also, EVT has been the object of many recent papers and developments. Such is the vastness of the field that the work here presented is, and could only be, by no means exhaustive, neither theoretically nor practically speaking. The analysis subjected to the Case Study data is but a sample of all the statistical methodologies we could use to infer on this subject.

Given this mind-set, and aiming to answer the previous questions, this dissertation is organized as follows: on Chapter 2 will be presented a summarized overview of the probability theory that supports the EVT and the statistical approaches that will be dicussed in Chapter 3, including both parametric and semi-parametric methodologies; Chapter 4 will consist on the application of the methods presented in Chapter 3 to a data base of apnea times of competitive female freedivers, from the discipline of static apnea; finally, in Chapter 5 there will be a discussion of the results obtained in Chapter 4, on an attempt to answer some of the questions that arise regarding the data, plus the presentation of some conclusions and possible further investigation to be done on the data with other existing methodologies that are beyond this dissertation.

Chapter 2

Probabilistic Overview of Extreme Value Theory

Many real life processes are only as relevant as their most differentiated occurrences, when a single event can, because of its magnitude, overbear the importance of all the other more central observations combined – for what they are called *extreme* events. For example, we concern ourselves with the maximum wind speed that would bring down a suspended bridge, not so much with the effect of the everyday air currents. But just as frequently the records of such events are scarce or nonexistent – for what they are called *rare* events.

These are the types of events with which the Extreme Value Theory is concerned. EVT comprises a probabilistic framework that supports several statistical procedures for dealing with extreme and rare events, that allow us to make predictions on certain parameters of interest, such as the most extremal value that a process can assume (*endpoint*), the probability that a specific threshold will be exceeded or the average amount of time separating two occurrences of that kind (*exceedance probabilities* and *return periods*).

To this end, ordering the sample is a vital step, i.e., representing the sample in the way that we clearly identify its *minimum* and/or *maximum* values. To the ordered observations of a sample we call order statistics, and the theoretical study of these quantities' properties is in itself a broad field of Probability theory – see, for example, Arnold et al. (2008). Still, to properly understand EVT, some knowledge of the exact and asymptotical distributional theory of o.s. is mandatory.

From here on out, let X be a continuous random variable with distribution function (d.f.) F and probability density function (p.d.f.) f, and $(X_1, X_2, ..., X_n)$ a sample of n r.v.'s independent and identically distributed (i.i.d.) to X. Some theory exists for cases when independency or equal distribution cannot be assumed, but we will focus primarily on i.i.d. samples. Also, there has been work developed regarding multivariate extremes, where the samples are formed by grouped observations from different r.v.'s. See Tiago de Oliveira and Gomes (1984) for an illustrative approach to bivariate models, for this topic falls outside the scope of this dissertation.

We represent the ordered sample as $(X_{1:n}, X_{2:n}, \cdots, X_{n:n})$, where the order defined is non

decreasing $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$. Therefore, $X_{i:n}$ represents the i$^{\text{th}}$ o.s., $i \in \{1, \cdots, n\}$, and the minimum and maximum of the sample correspond to $X_{1:n}$ and $X_{n:n}$, respectively. We will briefly approach some distributional results for these entities.

We can obtain the minimum of a sample simply by considering the relation $\min(X_1, \cdots, X_n) = -\max(-X_1, \cdots, -X_n)$. Thus, all the results regarding minima can be obtained from the results for the maxima. As such, in this dissertation all the results presented will be in the scope of maxima. In Reiss and Thomas (2007) and Coles (2001) some of the results for minima can be explicitly found, as well as the application of the aforementioned relation.

When we study a process in average terms, it is easy to recognize the importance of developing limiting laws for sums of increasing number of terms, such that the Central Limit Theorem (CLT) is one of the first asymptotic results lectured on introductory courses of Probability and Statistics. The CLT is of such relevance that further work on the result has been vastly pursued – see Fischer (2011) – and it is considered the unofficial sovereign of Classical Statistics. We will see ahead how a similar result is necessary, even essential to the development of EVT.

2.1 On the Sample Maximum and Other Order Statistics

As referred earlier, the study of order statistics plays a key role on EVT, for the extreme values embody *the best* and/or *the worst* in a sample and can only be found upon its assortment. Thus, it is only natural that the distributional properties of this quantities are the starting point for building the theoretical framework of EVT.

We will only be formulating some basic results on the exact distributions of o.s.'s. A simple enough method for obtaining all of the following expressions can be found in Castillo et al. (2004), which serves as proof to the next two theorems. The first refers to the distribution of a single o.s. and the second concerns the joint distribution of a set of order statistics from a sample. To relieve the notation, consider from this point $M_n := X_{n:n} = \max(X_1, ..., X_n)$ the sample maximum and $m_n := X_{1:n} = \min(X_1, ..., X_n)$, the sample minimum.

Theorem 2.1.1. *Let $(X_1, X_2, ..., X_n)$ be a sample of r.v.'s i.i.d. to X, with d.f. F and p.d.f. f. The d.f. and p.d.f. of the i^{th} order statistic $X_{i:n}$ are respectively*

$$F_{i:n}(x) = \sum_{r=i}^{n} \binom{n}{r} F^r(x)[1 - F(x)]^{n-r} \tag{2.1}$$

and

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!} F^{i-1}(x)[1 - F(x)]^{n-i} f(x) \, . \tag{2.2}$$

An alternative method for deriving this expressions to the one shown in Castillo et al. (2004) can be found on Arnold et al. (2008). Since we will be working with the sample's maximum or minimum, we can find the distributions functions of these particular o.s.'s by direct application of (2.1) – as done here to find the d.f. of m_n:

$$F_{1:n}(x) = 1 - [1 - F(x)]^n \; ; \tag{2.3}$$

or by following the definition of d.f. – as done here to find the d.f. of M_n:

$$F_{n:n}(x) := P\{M_n \leq x\} = P\{X_1 \leq x, \cdots, X_n \leq x\} = \prod_{i=1}^{n} P\{X_i \leq x\} = F^n(x) \ . \tag{2.4}$$

Note that both methods would wield the same expressions.

The specified probability distribution functions for m_n and M_n can be obtained by derivation of (2.3) and (2.4), resp., or directly form (2.2):

$$f_{n:n}(x) = nF^{n-1}(x)f(x) \ , \tag{2.5}$$

$$f_{1:n}(x) = n[1 - F(x)]^{n-1}f(x) \ . \tag{2.6}$$

We will also be interested in understanding how a group of o.s.'s from the same sample are jointly distributed.

Theorem 2.1.2. *Under the conditions of Theorem 2.1.1, let $X_{r_1:n}, ..., X_{r_k:n}$, $1 \leq r_1 < \ldots < r_k \leq n$, be a subset of k o.s. from that sample. The joint p.d.f. of this subset for order statistics is*

$$f_{r_1:n,...,r_k:n}(x_1, ..., x_k) = n! \prod_{i=1}^{k} f(x_i) \prod_{j=1}^{k+1} \frac{[F(x_j) - F(x_{j-1})]^{r_j - r_{j-1} - 1}}{(r_j - r_{j-1} - 1)!} \quad , x_1 \leq ... \leq x_k \ . \tag{2.7}$$

Applying (2.7) expressly to the largest k order statistics we obtain one more useful expression:

$$f_{(n-k+1):n,...,n:n}(x_{n-k+1}, ..., x_n) = n! \prod_{i=n-k+1}^{n} f(x_i) \frac{F(x_{n-k+1})^{n-k}}{(n-k)!} \quad , x_{n-k+1} \leq ... \leq x_n \ . \tag{2.8}$$

Similarly to what happens when we are dealing with sums of the r.v.'s in a sample, in the context of Classical Statistics, in the context of Extremes the exact distributional theory is not enough, specially when the distribution F is unknown. There is the need to study the asymptotic behaviour of the order statistics. We will consider three types of o.s.'s, which we can define as:

Central Order Statistics $X_{k:n}$ where $k = k_n \to \infty$ but $\frac{k}{n} \to \lambda$, $0 < \lambda < 1$ as $n \to \infty$;

Extremal Order Statistics $X_{k:n}$ (*lower extremal*) or $X_{n-k:n}$ (*upper extremal*) when k is a given fixed integer;

Intermediate Order Statistics $X_{k:n}$ where $k = k_n \to \infty$ but $\frac{k}{n} \to 0$ or $\frac{k}{n} \to 1$ as $n \to \infty$.

According to these definitions, we have that the sample maximum M_n is an upper extremal o.s., as its order is set as the sample size n. Analysing the behaviour of this statistic as the sample size increases towards infinity, we recognize that its limit distribution law is degenerated. Since

$$F^n(x) \underset{n\to\infty}{\longrightarrow} \begin{cases} 0, & F(x) < 1 \\ 1, & F(x) = 1 \end{cases} \ , \tag{2.9}$$

we have $M_n \xrightarrow[n\to\infty]{d} x^F$, where $x^F := \sup\{x : F(x) < 1\} \leq \infty$ is the right endpoint of the distribution F. The latter convergence in distribution of the maximum to a constant implies the convergence in probability $M_n \xrightarrow[n\to\infty]{p} x^F$ (in fact, since the o.s.'s are assorted in an ascending order, the stronger almost sure convergence $M_n \xrightarrow{a.s.} x^F$ also stands).

So the asymptotic distribution of the maximum is degenerated, and therefore useless for inference purposes. As such, some normalization is required to find a non-degenerated asymptotic distribution that will allow us to avoid this difficulty. We will look for an analogous result to the Central Limit Theorem when dealing with sums, for which we need to find convenient real sequences $a_n > 0$ and b_n. These will stabilize resp. the scale and location of M_n as the sample size n increases.

It is then possible to deal only with the linearly normalized maximum and find its asymptotic behaviour

$$\frac{M_n - b_n}{a_n} \xrightarrow[n\to\infty]{d} Y \backsim G \,. \tag{2.10}$$

The existence of appropriate choices for the normalizing sequences such that the limit distribution G is non-degenerate is a question with two layers. Beirlant et al. (2004) entitle them as the *extremal limit problem* and the *domain of attraction problem*, titles that were maintained in this dissertation. The extremal limit problem consists in finding all the possible limiting distributions that can appear in (2.10), whereas the domain of attraction problem dwells with the characterization of the distributions F for which these normalizing sequences exist and allow (2.10) to hold for any such specific limit distribution, as well as the specification of such sequences. These are the topics concerning the following sections.

2.2 Limiting Distributions for Maxima – The Extremal Limit Problem

The convergence in (2.10) is equivalent to

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \tag{2.11}$$

for all the continuity points of the distribution G. We need now to learn what forms can G take, so we start by defining the useful concept of *type*.

Definition 2.2.1. *Two d.f.'s U_1 and U_2 are of the* ***same type*** *if there exist real constants $A > 0$ and B for which $U_2(x) = U_1(Ax + B)$, for any real value of x.*

This also means that U_1 and U_2 belong to the same location-scale family (they are the same distribution, differentiated only by the location and scale parameters). The next theorem dictates the conditions under which this relation happens, known as **Convergence to Types Theorem** (of Khinchine). A proof can be found in Feller (1966).

Theorem 2.2.1. *Let U_1 and U_2 be two non-degenerate d.f.'s. If for F_n a sequence of d.f.'s there exist real sequences $a_n, \alpha_n > 0$ and b_n, β_n such that*

$$\lim_{n\to\infty} F_n(a_n x + b_n) = U_1(x) \text{ and } \lim_{n\to\infty} F_n(\alpha_n x + \beta_n) = U_2(x) \tag{2.12}$$

then exist $A > 0$ and B real for which

$$\frac{\alpha_n}{a_n} \underset{n\to\infty}{\longrightarrow} A \ , \ \frac{\beta_n - b_n}{a_n} \underset{n\to\infty}{\longrightarrow} B \tag{2.13}$$

and U_1 and U_2 are of the same type (according to Definition 2.2.1). Inversely, if (2.13) holds, then any of the relations in (2.12) implies the other one and the belonging to the same type of U_1 and U_2.

Another important concept is that of *max-stability*.

Definition 2.2.2. *A r.v. X with d.f. F is **max-stable** (or alternatively, F is a max-stable distribution) if, for any $n \in \mathbb{R}$, there are normalizing real constants $A_n > 0$ and B_n such that $F^n(x) = F(A_n x + B_n)$.*

This definition is saying that the distribution of a sample maximum $X_{n:n}$ is of the same type of the distribution of the r.v. X, thus the term stability of the maximum. If a limiting distribution for the normalized maximum exists, that distribution will have to be max-stable.

We are now able to enunciate the theorem that presents a unified version of the only possible non-degenerate asymptotic distributions for the normalized sample maximum. This is vastly accepted as the fundamental theorem of EVT, playing an analogous role to the CLT in Classical inference (of course with some differences), and it is the answer to the extremal limit problem. The version shown here is the result of the work of Fisher and Tippett (1928) and Gnedenko (1943) (for which it is also known as the **Fisher-Tippett-Gnedenko Theorem**), synthesized by a parametrisation presented by von Mises (1936) and Jenkinson (1955). A proof of the result, in the form presented here, can be found in de Haan and Ferreira (2006).

Theorem 2.2.2 (**Asymptotic Distribution of the Sample Maximum**). *If there exist real sequences $a_n > 0$ and b_n such that the limit in (2.11) is true for every continuity point x of G a non-degenerate distribution, then G is of the same type of*

$$G_\xi(x) := \begin{cases} \exp(-[1+\xi x]_+^{-1/\xi}), & \text{if } \xi \neq 0 \\ \exp(-\exp(-x)), & \text{if } \xi = 0 \end{cases} \tag{2.14}$$

for some value of $\xi \in \mathbb{R}$, where $x_+ := \max\{0, x\}$.

Note that, unlike the CLT, this theorem does not guarantee that such sequences exist, and they don't always do. The distribution G_ξ in (2.14) is known as the *Generalized Extreme Value Distribution* – or **GEV** – and its shape parameter ξ is denominated *Extreme Value Index* (EVI), a very important quantity in EVT, for it can be seen as a measure of heaviness for the tail of the underlying distribution F of X. In fact, the GEV is the only possible max-stable distribution,

and the condensation of the three standard distribution functions that can appear in (2.10), and which correspond to having a positive, negative or null EVI. As such, the type of G can be the same as:

Type I – *Gumbel*: if $\xi = 0$ in (2.14), we have

$$\Lambda(x) := \exp(-\exp(-x)) \equiv G_0(x) \ , \ x \in \mathbb{R} \tag{2.15}$$

(obtained as the continuity limit as $\xi \to 0$)

Type II – *Fréchet*: if $\xi > 0$ in (2.14), we have

$$\Phi_{1/\xi}(x) := \left\{ \begin{array}{ll} 0, & x < 0 \\ \exp(-x^{-1/\xi}), & x \geq 0 \end{array} \right\} \equiv G_\xi \left(\frac{x-1}{\xi} \right) \tag{2.16}$$

Type III – *Max-Weibull*: if $\xi < 0$ in (2.14), we have

$$\Psi_{-1/\xi}(x) := \left\{ \begin{array}{ll} \exp(-(-x)^{-1/\xi}), & x \leq 0, \\ 1, & x > 0 \end{array} \right\} \equiv G_\xi \left(\frac{x+1}{-\xi} \right) \tag{2.17}$$

These three types and the GEV itself can be expressed by their respective location/scale families, by introducing location and scale parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$:

$$V(x|\mu, \sigma) = V\left(\frac{x-\mu}{\sigma} \right), \quad \text{with } V = \Lambda,\, \Phi_{1/\xi},\, \Psi_{-1/\xi},\, G_\xi \,. \tag{2.18}$$

Even though we know Gumbel, Fréchet and Max-Weibull are the only possible limiting distributions for the normalized maximum, Theorem 2.2.2 does not allow a direct identification of which one type appears in the limit. It is then necessary to determine the signal of the EVI, or equivalently to which *domain of attraction* does the F distribution belong.

Definition 2.2.3. *Is is said that F belongs to the **max-domain of attraction** of G_ξ, and denoted $F \in \mathcal{D}_\mathcal{M}(G_\xi)$, if G in the limit (2.10) is of the same type as G_ξ. Particularly, if $\xi = 0$, $\xi > 0$ or $\xi < 0$, then F belongs to the Gumbel, Fréchet or Weibull domain of attraction, respectively.*

Definition 2.2.4. *We define the **right tail** of a distribution function F as $\overline{F}(x) := P(X > x) = 1 - F(x)$.*

The domain of attraction to which the distribution F belongs, determined by the signal of the EVI ξ, is intimately associated with the tail heaviness of the distribution, as mentioned earlier, and consequently with the right endpoint x^F. The EVI indicates the speed of decay to 0 of $\overline{F}(x)$ as x approaches the right endpoint x^F:

- $\xi < 0$ (Weibull domain) indicates a short tail, with finite right endpoint x^F;
- $\xi = 0$ (Gumbel domain) indicates an exponential tail (to which the others are compared, making it a changing point), with possibly finite or infinite right endpoint x^F;

- $\xi > 0$ (Fréchet domain) indicates a heavy tail (polynomial decay to 0) and the right endpoint x^F is infinite.

In the next section we will concern ourselves with the conditions that allow the identification of the max-domain of attraction for a specific distribution F.

2.3 Max-Domain of Attraction – The Domain of Attraction Problem

The problem of characterizing each max-domain of attraction is quite complex, but necessary. There are a number of conditions developed by several scientists that outline the set of distributions F for which the suitably normalized maximum converges to a given possible limiting distribution (presented in the previous section). Some of these conditions are very difficult to verify, and some don't guarantee equivalence, being only sufficient or necessary. In this dissertation it will only be presented three sets of noted conditions, with some of their variations, and a way of finding appropriate normalizing sequences for the sample maximum. All the results and respective proofs can be found in de Haan and Ferreira (2006).

Firstly, we need to define some concepts.

Definition 2.3.1. *Let F be a continuous d.f. with inverse $F^{\leftarrow}(y) = \inf\{x : F(x) \geq y\}$; we designate as* ***tail quantile function***

$$U(t) := F^{\leftarrow}\left(1 - \frac{1}{t}\right), \quad t \in [1, \infty[.$$

This function has the following properties:

- *$U(t)$ is monotonous and non-decreasing;*
- *$U(t)$ can be interpreted as the* return level *for the* return period *t;*
- *$U(1) = \inf\{x : F(x) \geq 0\}$ is the left endpoint x_F;*
- *$U(\infty) \equiv \lim_{t\to\infty} U(t) = \inf\{x : F(x) \geq 1\}$ is the right endpoint x^F.*

Definition 2.3.2. *If a positive function h is such that $\lim_{t\to\infty} \frac{h(tx)}{h(t)} = x^\beta$, for $x > 0$, it is said that h is* ***regular varying with index*** *β (at infinity) and denoted $h \in \mathcal{RV}_\beta$. Also, if a function a is satisfies $\lim_{t\to\infty} \frac{a(tx)}{a(t)} = 1$, for $x > 0$, then a is* ***slow varying*** *(at infinity) and denoted $a \in \mathcal{RV}_0$.*

Regarding these last concepts of *Regular and Slow Variation*, developed by de Haan (1970), we know that if a distribution is such that its tail function is slow varying (that is $\overline{F} \in \mathcal{RV}_0$), then it will not belong to any max-domain of attraction. This is called a super-heavy tail, and further material on such functions can be found in Fraga Alves et al. (2009a) and Fraga Alves et al. (2011).

We now present a set of relations rendered by Laurens de Haan that improve and are equivalent to the condition

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G_\xi(x) \quad \Leftrightarrow \quad F \in \mathcal{D}_{\mathcal{M}}(G_\xi),$$

for some real value of the EVI. The theorem in this dissertation is the reformulation of the original result, with the posterior knowledge of the GEV as the only possible limiting distribution for the normalized maximum.

Theorem 2.3.1 (**de Haan and Ferreira (2006) Theorem 1.1.6**). *For $\xi \in \mathbb{R}$ and G_ξ the GEV d.f. in* (2.14)*, the following statements are equivalent:*

1. *There exists real constants $a_n > 0$ and b_n such that*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G_\xi(x) \tag{2.19}$$

   for all x continuity point of G_ξ;

2. *There is a positive function a such that for $x > 0$*

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\xi - 1}{\xi} \tag{2.20}$$

   where the right hand side is interpreted as $\log(x)$ for $\xi = 0$;

3. *There is a positive function a such that*

$$\lim_{t\to\infty} t\,(1 - F\,(a(t)x + U(t))) = (1 + \xi x)^{-1/\xi} \tag{2.21}$$

   for all x with $1 + \xi x > 0$;

4. *There exists a positive function f such that*

$$\lim_{t\uparrow x^F} \frac{1 - F\,(t + x f(t))}{1 - F(t)} = (1 + \xi x)^{-1/\xi} \tag{2.22}$$

   for all x for which $1 + \xi x > 0$.

Moreover, (2.19) *holds with $b_n := U(n)$ and $a_n := a(n)$. Also,* (2.22) *holds with $f(t) = a\left(\frac{1}{1-F(t)}\right)$.*

The expression (2.20) in the second statement of this theorem corresponds to the definition of *Extended Regular Variation* (see Appendix B in de Haan and Ferreira (2006)), and so this condition is known as First Order Extended Regular Variation Property, usually referred simply as **First Order Condition**. This is a necessary and sufficient condition for characterizing the max-domains of attraction, very useful in proving other conditions with the same objective.

The Second Order Extended Regular Variation Property, usually referred simply as **Second Order Condition** as we now will enunciate, provides us with further information on the tail of the distribution function F, specifying the speed of convergence in the First Order Condition, that

is, the speed at which the normalized maximum's d.f. approximates the limiting distribution G_ξ. Let us then consider the First Order Condition with the simpler notation for the limit function

$$D_\xi(x) := \begin{cases} \dfrac{x^\xi - 1}{\xi}, & \xi \neq 0 \\ \log(x), & \xi = 0 \end{cases}.$$

According to de Haan and Stadtmüller (1996), we assume the existence of a *second order auxiliary function* A positive or negative (but of unchanging sign) such that $\lim_{t\to\infty} A(t) = 0$ and, under the conditions of the First Order Condition,

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - D_\xi(x)}{A(t)} = H_{\xi,\rho}(x) := \frac{1}{\rho}\left(\frac{x^{\xi+\rho}-1}{\xi+\rho} - \frac{x^\xi - 1}{\xi}\right), \tag{2.23}$$

for all $x > 0$ and with the *second order parameter* $\rho \leq 0$. Furthermore, the A function is such that $|A| \in \mathcal{RV}_\rho$. Continuity arguments are used in considering the limit of $H_{\xi,\rho}$ when $\xi = 0$ or $\rho = 0$. Let us then formalize this concept in the following definition of the Second Order Condition.

Definition 2.3.3. *A function U or its associated d.f. F is said to satisfy the* ***Second Order Condition*** *if, for some positive function a and for some positive or negative function A, with $\lim_{t\to\infty} A(t) = 0$, condition (2.23) is verified.*

As such, the function A describes the convergence rate of the first order condition for a d.f. F, determined by its index ρ of regular variation: negative values of the second order parameter show a fast (algebraic) speed of convergence in (2.20), while $\rho = 0$ determines a slower speed of convergence of the normalized M_n, for instance at a logarithmic rate. Note that the Second Order Condition implies the First Order one, and consequently the belonging of F to a max-domain of attraction. It has been proved by Dekkers and de Haan (1989) that this holds for most well known distributions, such as the Normal, Exponential and GEV distributions.

Years before de Haan disclosed the equivalent conditions mentioned above, von Mises (1936) came up with a set of sufficient conditions to characterize d.f.'s attracted to the Weibull, Gumbel and Fréchet max-domains. These separate conditions were then synthesized in one general formulation known as **von Mises' Sufficient Condition**, which will be enunciated in the following theorem, with a variation in terms of the tail quantile function.

Theorem 2.3.2. *For a continuous d.f. F suppose the p.d.f. $f(x) := F'(x)$ and $F''(x)$ exist; define the* hazard function *and its algebraic inverse respectively by*

$$h(x) := \frac{f(x)}{1-F(x)} \quad \textit{and} \quad r(x) := \frac{1-F(x)}{f(x)}.$$

If

$$\lim_{x\to x^F} r'(x) = \xi \tag{2.24}$$

then $F \in \mathcal{D}_\mathcal{M}(G_\xi)$ with normalizing constants $b_n = U(n)$ and $a_n = nU'(n)$. Also, condition 2.24

is equivalent, in terms of U, to

$$\lim_{t\to\infty} \frac{tU''(t)}{U'(t)} = \xi - 1 \tag{2.25}$$

which by Theorem 2.3.1 implies that $F \in \mathcal{D}_{\mathcal{M}}(G_\xi)$.

Simpler von Mises' conditions are possible for $\xi \neq 0$ either in terms of the F function, or in terms of the U function, and shown in de Haan and Ferreira (2006).

Later, Gnedenko (1943) introduced new simple necessary and sufficient conditions for maximal attraction to the three types of limit laws.

Theorem 2.3.3 (**Gnedenko's Necessary and Sufficient Conditions**). *The distribution function F is in the domain of attraction of the extreme value distribution G_ξ if and only if*

1. *for $\xi > 0$: x^F is infinite and $\overline{F} \in \mathcal{RV}_{-1/\xi}$, i.e.,*

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\xi}, \quad \forall x > 0; \tag{2.26}$$

2. *for $\xi < 0$: x^F is finite and*

$$\lim_{t\downarrow 0} \frac{1 - F(x^F - tx)}{1 - F(x^F - t)} = x^{-1/\xi}, \quad \forall x > 0; \tag{2.27}$$

3. *for $\xi = 0$: x^F can be finite or infinite and*

$$\lim_{t\uparrow x^F} \frac{1 - F(t + xf(t))}{1 - F(t)} = e^{-x}, \quad \forall x \tag{2.28}$$

where f is a suitable positive function. If (2.28) holds for some f, then $\int_t^{x^F} (1 - F(s))\, ds < \infty$ for $t < x^F$ and it holds with f the Mean Excess Function, *defined as*

$$f(t) := \frac{\int_t^{x^F} (1 - F(s))\, ds}{1 - F(t)} = E[X - t | X > t], \quad for\ t < x^F. \tag{2.29}$$

These conditions (that, once again, do not grant the existence of a limiting distribution for the sequence of suitably normalized maxima) were later unified in a seemingly more uniform way by de Haan. In statement *4.* of Theorem 2.3.1 a suitable positive function f is mentioned, but not presented. de Haan showed that this function depends on the signal of the EVI parameter and used the referred condition as a definition of belonging to a max-domain of attraction to provide an alternative version of the necessary and sufficient conditions of Gnedenko.

Theorem 2.3.4 (**de Haan and Ferreira (2006) Theorem 1.2.5**). *The distribution function F is in the domain of attraction of the extreme value distribution G_ξ if and only if for some positive function f*

$$\lim_{t\uparrow x^F} \frac{1 - F\left(t + xf(t)\right)}{1 - F(t)} = (1 + \xi x)^{-1/\xi} \tag{2.30}$$

for all x for which $1+\xi x > 0$. If (2.30) holds for some $f > 0$ then it also holds for

$$f(t) = \begin{cases} \xi t, & \xi > 0 \\ -\xi(x^F - t), & \xi < 0 \\ \dfrac{\int_t^{x^F} (1 - F(s))\, ds}{1 - F(t)}, & \xi = 0. \end{cases} \tag{2.31}$$

Also by Gnedenko, a suggestion on how to find possible normalizing constants in the basic limit relation (2.11) was explored that consists in the most common choice for a_n and b_n. As we know by the Convergence of Types Theorem (2.2.1) this choice is not unique, and even possibly differs from the ones already presented, for instance, in Theorem 2.3.2, for the von Mises' condition. The constants here presented are dependant on the type of function G that figures in (2.11).

Theorem 2.3.5 (**Normalizing Constants**). *Suppose $F \in \mathcal{D}_{\mathcal{M}}(G_\xi)$; then*

1. *for $\xi > 0$:*
$$\lim_{n\to\infty} F^n(a_n x) = \exp\left(-x^{-1/\xi}\right) = \Phi_{1/\xi}(x),$$
for $x > 0$, with $a_n = U(n)$ and $b_n = 0$;

2. *for $\xi < 0$:*
$$\lim_{n\to\infty} F^n(a_n x + x^F) = \exp\left(-(-x)^{-1/\xi}\right) = \Psi_{-1/\xi}(x),$$
for $x > 0$, with $a_n = x^F - U(n)$ and $b_n = x^F$;

3. *for $\xi = 0$:*
$$\lim_{n\to\infty} F^n(a_n x + b_n) = \exp\left(-e^{-x}\right) = \Lambda(x),$$
for all x, with $a_n = f\left(U(n)\right)$, $b_n = U(n)$ and f the Mean Excess Function as in Theorem 2.3.3.

2.4 Limiting Distributions for Excesses

Another way to look at the extreme events in a sample is to position ourselves in a certain value – *threshold* – and analyze all the observations that fall beyond that level – *exceedances*. This is an appropriate approach for modeling the extremes of data when a complete series is available, since there is less waste of information when comparing to modeling based solely on the sample maximum.

The assumed conditions about the model are the same made in the introduction to this Chapter, with special emphasis on the i.i.d. nature of the random variables. So, given a r.v. X with continuous d.f. F, our focus is on characterizing the conditional distribution and asymptotic behaviour of the *excesses* over a given threshold – differences between the values of the exceedances and the threshold itself. If we define the random variable $Y := X - u$, where u is

our threshold, then $Y|Y>0$ represents the r.v. of the excesses conditional to the exceedances and its distribution function F_u can be given by

$$F_u(y) := P[X-u \leq y | X > u] = \frac{F(y+u)-F(u)}{1-F(u)}, \quad 0 \leq y \leq x^F - u\,. \tag{2.32}$$

Therefore, if the "parent" distribution F is given we can directly obtain the distribution of the threshold exceedances. This is seldom the case in practice, and analogously to what happens when modeling sample maxima resorting to the GEV, limiting approximations to F_u are sought, with the particularity that such approximation has to hold for a vast choice of suitable thresholds.

Let us introduce the *Generalized Pareto Distibution* – or **GP** – which we denote by H_ξ and is defined as

$$H_\xi(y|\sigma_u) := \begin{cases} 1-\left(1+\dfrac{\xi y}{\sigma_u}\right)^{-1/\xi}, & y \in (0;\infty),\ \xi > 0 \\ 1-\exp\left(-\dfrac{y}{\sigma_u}\right), & y \in (0;\infty),\ \xi = 0 \\ 1-\left(1+\dfrac{\xi y}{\sigma_u}\right)^{-1/\xi}, & y \in (0;-\dfrac{\sigma_u}{\xi}),\ \xi < 0 \end{cases} \tag{2.33}$$

where, as before, the shape parameter ξ is the EVI, and σ_u is a scale parameter, already here indexed in u because of its relation to the threshold of the excesses model, as will be shown ahead. Note that, for $\xi = 0$, $\xi > 0$ and $\xi < 0$ respectively, H_ξ is reduced to the Exponential, type II Pareto and Beta distributions. A reparameterization in x is possible simply by considering $x = y + u$, adding a location parameter u to the distribution (see, for example, Fraga Alves and Neves (2015)).

Pickands (1975) and Balkema and de Haan (1974) proposed an approximation for F_u, when the value of u is suitably high, to the GP. This allowed for a duality to be established between the GEV of shape ξ and the GP with the same shape parameter ξ. Thus, let us now enunciate the theorem that determines the approximation

$$F_u(y) \approx H_\xi(y|\sigma_u), \quad y > 0, \tag{2.34}$$

known as **Pickands-Balkema-de Haan Theorem**, or the second theorem of EVT.

Theorem 2.4.1.

$$F \in \mathcal{D}_{\mathcal{M}}(G_\xi),\ \xi \in \mathbb{R} \quad \Leftrightarrow \quad \lim_{u \to x^F} \sup_{0<y<x^F-u} |F_u(y) - H_\xi(y|\sigma_u)| = 0$$

where σ_u is used to imply that the scale parameter depends on the threshold u.

This theorem implies that if an approximation of the sample maxima in the GEV family is appropriate, then the threshold excesses have a corresponding limiting distribution within the GP family, with its parameters determined by those associated with the GEV. Thereby, the EVI is to the GP distribution as dominant in determining the tail behaviour as it is for the GEV: F_u has a heavy upper tail with infinite upper limit for $\xi > 0$, a light upper tail with finite right

endpoint given by $u - \frac{\sigma_u}{\xi}$ for $\xi < 0$ and an exponential right tail for $\xi = 0$; this behaviour of course extends to the upper tail of F.

Note that the result does not directly depend on the value of the threshold, i.e., for a variety of conveniently hight values of u the value of the EVI should be the same, meaning that the max-domain of attraction of F remains the same for all such choices of u, even though this influences the scale parameter. It can also be said that F belongs to the POT-domain of attraction of H_ξ, although this denomination is less used (see Fraga Alves et al. (2011)).

In Coles (2001) an outline justification for the approximation to the Generalized Pareto model can be found.

Regarding this approach remains the issue of what exactly is a suitably hight threshold. This question is very hard to answer and there is some controversy surrounding it. As we have seen, it seems that the choice of threshold is indifferent to the determination of the domain of attraction to which a distribution belongs. However, we will later see that, statistically, this is not true, and that the choice of value for u must be made carefully. This is easily understood if we think that a very low threshold leads us to consider events that aren't extreme at all, or that a too high threshold can lead us to work solely with the sample maximum, for example. So, this topic must be object of careful consideration.

2.5 Limiting Distributions for the Largest Order Statistics

As stated before, the sample maximum M_n is an upper extremal order statistic, and limiting our study to this quantity can mean loosing information provided by the sample, especially considering extremes are, by definition, scarce. In the last section, we introduced a way of dealing with this issue, by considering all the observations above a certain level. We will know present a different approach that has the same objective of using more of the sample's information, based on the behaviour of the largest k o.s.'s.

Let us denote here $M_n^{(k)} := X_{n-k+1:n}$ the k^{th} *largest o.s.*, with k a fixed integer. We can find in Gomes et al. (2013a) a justification for the validity of the following result regarding the asymptotic distribution of $M_n^{(k)}$:

Theorem 2.5.1. *For a fixed integer k and $M_n^{(k)}$ as defined previously, we have*

$$\lim_{n\to\infty} P[M_n \leq a_n x + b_n] = G_\xi(x)$$

$$\Leftrightarrow$$

$$\lim_{n\to\infty} P[M_n^{(k)} \leq a_n x + b_n] = G_\xi(x) \sum_{i=0}^{k-1} \frac{[-\log(G_\xi(x))]^i}{i!} \tag{2.35}$$

with $\xi \in \mathbb{R}$ and constants $a_n > 0$ and $b_n \in \mathbb{R}$.

It should be noted here that the normalizing constants for $M_n^{(k)}$ are the same (or at least asymptotically equivalent in the sense of Theorem 2.2.1) as the ones for the sample maximum.

Also, the d.f. $G_\xi(x)$ in (2.35) is the same to which max-domain of attraction F belongs, being its parameters defined univocally by the limiting GEV distribution of the sample maximum. The limiting probability distribution function of the k^{th} largest o.s. is then easily obtained by derivation of the limiting distribution function.

However, these o.s.'s are clearly not independent from each other, for different values of k, thus the outcome of each limit influences the distribution of the other, so Theorem 2.5.1 does not lead to a joint model for the largest k o.s.'s in itself, which we might be interested in studying as a way to better make use of the data at our disposal. This can be achieved, but the joint d.f. it leads to is intractable, for what we are only able to explicitly show the joint p.d.f.. A demonstration of the following result, regarding the asymptotic distribution of the k largest order statistics in a sample, can be found in Coles (2001).

Theorem 2.5.2. *For a fixed integer k and $M_n^{(k)}$ as defined previously, we have*

$$F \in \mathcal{D}_{\mathcal{M}}(G_\xi)$$

with $\xi \in \mathbb{R}$, constants $a_n > 0$ and $b_n \in \mathbb{R}$ and $g = G'_\xi$ if and only if the k-vector

$$\left(\frac{M_n^{(1)} - b_n}{a_n}, \ldots, \frac{M_n^{(k)} - b_n}{a_n} \right)$$

has a non-degenerate limiting distribution with joint p.d.f.

$$g_{1,\ldots,k}(w_1, \ldots, w_k) := G_\xi(w_k) \prod_{i=1}^{k} \frac{g(w_i)}{G_\xi(w_i)} \quad for \;\; w_1 > \ldots > w_k. \tag{2.36}$$

Note that if the specified value of k is 1, then we are reduced to the sample maxima case and to the GEV family of density functions. This is an indication of the suitability of this model.

As always, although our focus here is on the upper portion of the sample, equivalent results for the lower tail exist and can be found in the same bibliographic references already cited.

Chapter 3

Statistical Treatment of Extreme Value Data

Statistical inference in the Extreme Values field is of vital importance as, given the nature of the data it concerns, small estimation or preparation mistakes can have extremely severe consequences. In fact, statistical analysis of extreme events is the key to a safer management of risk situations in many fields, such as in Finance, Hydrology, Geophysics, Environment to enumerate only a few.

The EVI ξ estimation plays here a fundamental part, for it allows us to position ourselves in a certain domain of attraction. From there we can possibly obtain more accurate information on the location μ and scale σ parameters, as well as other interesting quantities like the right endpoint of the distribution, extreme quantiles, exceedance probabilities of high levels or return periods.

Given the scarce nature of this type of data and the consequent recurring necessity of extrapolating beyond the sample, some criticism has arisen stating that, even with the support of the asymptotic arguments exposed on the last Chapter, inferring on events never before registered is a leap of faith. But, in truth, such inference is indeed necessary and it is obviously preferable to do it resorting to techniques based on a rational theoretical base. This solution has been well accepted, since Extreme Value Analysis (EVA) has been proven to be a powerful and effective tool for providing information on atypical situations that may have significant impact. The biggest discord regards, then, the choice of methodology to apply.

We will here present two groups of approaches that are separated by the assumptions made *a priori* for the data and its underlying distribution. The first type consists on **parametric methods**, which are based on the assumption of a parametric extreme value model (such as the GEV or GP distributions) underlying the data. The second group regards **semi-parametric methods**, based not on the assumption of an underlying model, but solely on the belonging of its distribution function to a max-domain of attraction, for some real EVI.

3.1 Parametric Approach

We have justified previously how it is the behaviour of the right tail of the d.f. F underlying to the data that determines the max-domain of attraction, and a correct positioning in a specific type of max-domain leads to better and more accurate parameter estimation. Thus, we shine the spotlight on the higher order statistics, in detriment of the rest of the data available on the process at study. It is reasonable, then, to consider a parametric model from the GEV family of distributions to fit the sample of i.i.d. maxima observations, i.e., to use the limiting distribution of the sample extremes (sample maxima, in this dissertation) as an exact distribution to be fitted to the data.

Depending on the type of observations we consider, or more specifically, on how many top observations we base our inference on, different methodologies become pertinent, distinguished by the amount of information they drain from the data. Here will be presented three of the most mainstream methodologies:

The Gumbel Method also known as the Block (or Annual) Maxima method – BM – or classical (extremal) method, it is of the three here presented the one which uses less information from the sample, for it is only based on the single largest observation of each one of several blocks or time periods;

The Peaks Over Threshold Method or POT method, uses all the observations registered that fall beyond a suitably high and fixed threshold;

The Largest Observations Method or LO method, of which the Gumbel approach is a particular case (when considering $k = 1$), infers based on a fitted joint model for the largest k observations of each block or time period, with k a relatively small integer.

There will also be a brief reference to statistical inference when the assumption of i.i.d. observations is not reasonable, specifically when dealing with non-stationarity – the presence of structure in the series through time. But, unlike the previously mentioned approaches, there is still no unified theory on how to handle this kind of data, and as such it will only be presented a possible way of taking on this problem.

3.1.1 Extremal Classical Inference – Gumbel Method

The univariate GEV method, or Gumbel method named after it's developer, was the first model to be cultivated for Extremal Statistics, based on the limit result in Theorem 2.2.2. It consists of considering the sample of size n to be divided in m sub-samples, known as *blocks*, traditionally of equal dimension k such that $n = m \times k$, and fitting the GEV distribution to the sample of the m block maxima. Thus, let us consider the r.v. defined as $Y \equiv M_k = \max(X_1, \ldots, X_k)$ such that our maxima sample will be $(Y_1, ..., Y_m)$, to which the parametric model will be fitted. These are supposed to be independent variables, from a GEV distribution whose parameters will be estimated, since X_i are also assumed independent, although the independence of Y_i is probably still a reasonable approximation even if there is some dependance between the X_i variables.

In applying this model to actual data, the choice of the block size can be challenging, when not obvious from the nature of the variable. If the blocks are too small, the approximation to the limit model is likely to be unsatisfying, generating biased estimations, but too large blocks lead to a smaller maxima sample, producing heavy variability in the estimations, so the choice made must ideally be a compromise. Often, for practicality purposes, blocks are consider to represent one year of observations, meaning that the assumption that the block maxima have the same distribution is plausible. In many cases, this choice is suggested by the nature of the data, when for example only yearly maximal observations are available.

As previously stated, the estimation of the EVI is of great importance, and it is in this approach estimated jointly with the scale and location parameters. But the estimation process for the GEV distribution isn't always easy, so it is usual to do some preliminary fitting of the data to one of the particular extremal models (Gumbel (2.15), Fréchet (2.16) or Weibull (2.17)) and statistically choose the type best suited for the process at hand.

A common way to do this preliminary analysis if graphically, through probability or quantile plots (*qq-plots*). If the underlying model of the data is from the Gumbel family, there should be a linear relation between the empirical quantiles $y_{i:m}$ (sorted observations) and the theoretical Gumbel quantiles $-\log(-\log(p_i))$, with $p_i := {}^{i}/_{(m+1)}$ a possible definition of the plotting positions. This possible relation gives informal validation of the suitability of the Gumbel model and even preliminary estimates for the location μ and scale σ parameters, by the Least Squares method. Moreover, if the curve of said plot is approximately convex, it gives informal validation of the suitability of the Fréchet model; if, on the other hand, the curve appears to be concave, it gives informal validation of the suitability of the Weibull model. More on the preliminary approach can be found in Gomes et al. (2013a). Ahead it will be shown how to statistically choose a domain of attraction in a more defined way, resorting to hypothesis tests.

Many techniques for parameter estimation in Extremes exist in the literature, but the most attractive, due to their broad adaptability and simplicity, are the ones hereby presented: the *Maximum Likelihood* (ML) method and the *Probability Weighted Moments* (PWM) method.

3.1.1.1 Maximum Likelihood Estimation

When it comes to estimation, the most commonly used technique is the Maximum Likelihood method. Although it demands some significant computational effort, there exist nowadays several tools that make the process relatively simple, given the power of the current computers. This method, with little modification, allows us to deal with a variety of problems that can arise, such as non-stationarity or missing values in data.

So let $(y_1, \ldots, y_m)$ be an observed random sample of the r.v. Y defined above from the GEV family given by (2.14) and (2.18). The log-likelihood function for this sample is

$$\log L(\xi, \mu, \sigma | y_1, \ldots, y_m) = -m \log(\sigma) - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{m} \log\left(1 + \xi \frac{y_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \left(1 + \xi \frac{y_i - \mu}{\sigma}\right)^{-\frac{1}{\xi}} \tag{3.1}$$

for values of the EVI $\xi \neq 0$ and $1 + \xi \frac{y_i - \mu}{\sigma} > 0$, $i = 1, \ldots, m$. For samples associated with a EVI $\xi = 0$, the log-likelihood expression is

$$\log L(0, \mu, \sigma | y_1, \ldots, y_m) = -m \log(\sigma) - \sum_{i=1}^{m} \exp\left(-\frac{y_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \frac{y_i - \mu}{\sigma}. \tag{3.2}$$

Maximization of this equations with respect to the parameters (ξ, μ, σ) leads to the ML estimators $(\hat{\xi}, \hat{\mu}, \hat{\sigma})$ for the GEV family. Note that the optimization must be done through numerical algorithms, since no analytical solution exists – see, for example Smith (1985). As such, in some cases, lack of convergence of this methods can be a problem.

Much of the attractiveness of this method is tied to the useful properties of the estimators it yields. But for this asymptotic properties to stand, certain regularity conditions must be verified. This is a difficulty when working with the GEV family since the support of this distribution depends on its (unknown) parameters. It has been shown that consistency and asymptotic normality of the ML estimators depend on the value of the EVI: firstly, Smith (1985) proved that for $\xi > -0.5$ the ML estimators are consistent and

$$\sqrt{m}\left((\hat{\xi}, \hat{\mu}, \hat{\sigma}) - (\xi, \mu, \sigma)\right) \xrightarrow[m \to \infty]{d} Z \backsim \mathcal{N}(0, \mathcal{I}^{-1})$$

with $\mathcal{I}^{-1}$ being the inverse of the Fisher Information matrix; later Zhou (2009) and (2010) proved that this properties hold for $\xi > -1$.

For $\xi < -1$ the log-likelihood function has no local maximum and as such the procedure is not applicable. This fact is of small concern since this case corresponds to extremely short bounded upper tailed distributions, seldom encountered in real extreme value data sets. More details on the computational approach to the ML estimation, like the attainment of $\mathcal{I}^{-1}$, can be found in Castillo et al. (2004) and Beirlant et al. (2004). We will also see ahead how we can produce interval estimators from this theoretical framework.

3.1.1.2 Probability Weighted Moments Estimation

Another commonly used estimation procedure is the Probability Weighted Moments method, which consists of a generalization of the usual Moments method through the Probability Weighted Moments of a r.v. Y with d.f. F, presented by Greenwood et al. (1979),

$$M_{p,r,s} = E\left\{Y^p [F(Y)]^r [1 - F(Y)]^s\right\}, \quad p, r, s \in \mathbb{R}. \tag{3.3}$$

Its application to the GEV distribution has been extensively studied by Hosking et al. (1985). This method performs better than the ML method when working with small samples (which is not uncommon in EVA) as its estimators have lower variance than that of the ML estimators in such cases. However, the PWM method does not deal with structural problems in data as easily as the previous approach, being very difficult to modify the estimators to suit cases of non-stationarity, for example.

Being $(Y_1, \ldots, Y_m)$ the random sample of i.i.d. variables from the GEV family, the method uses the PWM for $p = 1$, $r = 0, 1, 2, \cdots$ and $s = 0$, given by

$$M_{1,r,0} = E\left\{Y[F(Y)]^r\right\} = \begin{cases} \dfrac{1}{r+1}[\mu + \sigma\{\epsilon + \log(1+r)\}], & \xi = 0 \\ \dfrac{1}{r+1}\left[\mu - \dfrac{\sigma}{\xi}\left\{1 - (r+1)^{\xi}\Gamma(1-\xi)\right\}\right], & \xi \neq 0,\ \xi < 1 \end{cases} \tag{3.4}$$

where $\epsilon = 0.57721$ is the *Euler's constant* and $\Gamma(.)$ the mathematical *gamma function*. The fact that such moments only exist for $\xi < 1$ is not problematic since in most real applications, values of the EVI usually fall between -0.5 and 0.5. Landwehr et al. (1979) proved that unbiased estimators for $M_{1,r,0}$ are given by

$$\hat{M}_{1,r,0} = \frac{1}{m}\sum_{i=1}^{m}\left(\prod_{k=1}^{r}\frac{(i-k)}{(m-k)}\right)Y_{i:m}. \tag{3.5}$$

Thus, solving the moments equation system as usual in order to (μ, σ) or (ξ, μ, σ) respectively if $\xi = 0$ or $\xi \neq 0$, using the corresponding expressions from (3.4), and replacing the PWM with their unbiased estimators in (3.5), we obtain the following PWM estimatores:

- if $\xi = 0$:

$$\hat{\sigma} = \frac{2\hat{M}_{1,1,0} - \hat{M}_{1,0,0}}{\log 2} \qquad \text{and} \qquad \hat{\mu} = \hat{M}_{1,0,0} - \epsilon\,\hat{\sigma}\ ; \tag{3.6}$$

- if $\xi \neq 0$:

$$\begin{aligned} &\hat{\sigma} = \frac{\hat{\xi}(2\hat{M}_{1,1,0} - \hat{M}_{1,0,0})}{\Gamma(1-\hat{\xi})(2^{\hat{\xi}} - 1)}, \qquad \hat{\mu} = \hat{M}_{1,0,0} + \frac{\hat{\sigma}}{\hat{\xi}}\left(1 - \Gamma(1-\hat{\xi})\right), \\ &\text{and } \hat{\xi} \text{ obtained numerically from} \qquad \frac{3\hat{M}_{1,2,0} - \hat{M}_{1,0,0}}{2\hat{M}_{1,1,0} - \hat{M}_{1,0,0}} = \frac{3^{\hat{\xi}} - 1}{2^{\hat{\xi}} - 1}\ . \end{aligned} \tag{3.7}$$

A more detailed derivation of these estimators can be found in Vicente (2012).

It was shown by Hosking et al. (1985) that for these PWM estimators, when $\xi < 1$ and $m \to \infty$,

$$\sqrt{m}\left((\hat{\xi}, \hat{\mu}, \hat{\sigma}) - (\xi, \mu, \sigma)\right)$$

is asymptotically Normal with a null mean vector. Details of this property are given in Beirlant et al. (2004).

3.1.1.3 Estimation of Other Relevant Indicators

Having estimated the core parameters of the GEV distribution as $(\hat{\xi}, \hat{\mu}, \hat{\sigma})$ through one of the methods indicated previously, we can use these estimates to infer on other interesting indicators such as *exceedance probabilities*, *return periods*, *return levels*, *extremal quantiles* and *the right endpoint*.

Exceedance Probability It is simply the probability that a (high) value u will be transcended, and it can be given directly by the tail function of the estimated GEV distribution:

$$\widehat{P(Y>u)} = 1 - G_{\hat{\xi}}(u|\hat{\mu},\hat{\sigma}) = \begin{cases} 1 - \exp\left\{-\left[1+\hat{\xi}\left(\frac{u-\hat{\mu}}{\hat{\sigma}}\right)\right]^{-1/\hat{\xi}}\right\}, & \hat{\xi}\neq 0 \\ 1 - \exp\left\{-\exp\left[-\frac{u-\hat{\mu}}{\hat{\sigma}}\right]\right\}, & \hat{\xi}=0 \end{cases}; \quad (3.8)$$

Return Period The return period of a level u is the average "amount of time" (number of blocks) it will take until a value larger than u is registered; it is deeply related to the concept of return level, and can be estimated by

$$\widehat{T(u)} = \frac{1}{\widehat{P(Y>u)}}; \quad (3.9)$$

Return Level This is defined by the tail quantile function in 2.3.1: it is the level exceeded on average once every t blocks, and can be estimated by inversion of the estimated GEV d.f.:

$$\widehat{U(t)} = G^{\leftarrow}_{\hat{\xi}}\left(1-\frac{1}{t}|\hat{\mu},\hat{\sigma}\right) = \begin{cases} \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}\left[(-\log(1-p))^{-\hat{\xi}} - 1\right], & \hat{\xi}\neq 0 \\ \hat{\mu} - \hat{\sigma}\,\log(-\log(1-p)), & \hat{\xi}=0 \end{cases} \quad (3.10)$$

with $p = 1/t$; given $U(t)$ the return level of t, we have $T(U(t)) = t$ the return period of $U(t)$;

Extremal Quantile This is the value that will be exceeded with a very small probability p, and it can be estimated in terms of (3.10) by:

$$\widehat{\chi_p} := G^{\leftarrow}_{\hat{\xi}}(1-p|\hat{\mu},\hat{\sigma}) = \widehat{U\left(\frac{1}{p}\right)}; \quad (3.11)$$

Right Endpoint In case the EVI is negative, $\xi < 0$, then we can estimate the value of the right endpoint as being the extremal quantile of probability 0:

$$\widehat{x^F} = \widehat{\chi_0} = \widehat{U(\infty)} = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}. \quad (3.12)$$

For estimating similar parameters for the underlying population X one simply needs to use the property of max-stability in 2.2.2. The deduction and its results can be found in Gomes et al. (2013a) or Vicente (2012).

3.1.1.4 Interval Estimation

As can be seen in Beirlant et al. (2004), confidence intervals (CI's) for the GEV parameters (ξ,μ,σ) can be derived from the asymptotic normality of its ML or PWM estimators. As such, these CI's are centered around the corresponding point estimates, given the symmetry of the

Normal distribution. This is not always on our best interest, and it has been shown that better interval estimations can be produced based on the *profile log-likelihood function*, which are not necessarily centered around the ML point estimate. This has been suggested by Beirlant et al. (2004) and more details can be found it said reference.

Definition 3.1.1. *For every value of ξ, the* ***profile log-likelihood function****, denoted $\log \mathcal{L}_p(\xi)$ gives the maximized log-likelihood function in order to the other parameters (μ, σ):*

$$\log \mathcal{L}_p(\xi) = \max_{(\mu,\sigma)|\xi} \{\log L(\xi, \mu, \sigma)\}.$$

The CI's are constructed from the *likelihood ratio test statistic*, defined as

$$\Lambda = \frac{\mathcal{L}_p(\xi_0)}{\mathcal{L}_p(\hat{\xi})},$$

which is used in testing the null hypothesis $H_0 : \xi = \xi_0$ against the corresponding alternative $H_1 : \xi \neq \xi_0$. Under H_0, we have

$$-2 \log \Lambda \xrightarrow[m\to\infty]{d} W \backsim \chi^2_1$$

and as such the rejection region of a test at the asymptotic level α is given by the condition $-2 \log \Lambda > \chi^2_1(1-\alpha)$, being the corresponding $(1-\alpha) \times 100\%$ confidence interval for the EVI derived as

$$CI_\xi = \left\{ \xi : \log \mathcal{L}_p(\xi) \geq \log \mathcal{L}_p(\hat{\xi}) - \frac{\chi^2_1(1-\alpha)}{2} \right\}, \tag{3.13}$$

With analogous reasoning, CI's can be constructed for the location and scale parameters, as well as for the other relevant indicators presented in the previous section, although a reparameterization of the GEV model might be necessary. This is easily done by considering the relations presented between these quantities of interest and the core parameters of the distribution, and thereby the presentation of the calculations is here excused.

3.1.1.5 Statistical Choice of Extreme Value Domains of Attraction

As stated before, it is important to choose the extremal model that most conveniently describes the d.f. of the population from which the data originated. For instance, wrongly selecting the Gumbel domain leads us to estimate the other parameters based on wrong likelihood expressions and consider the EVI fixed at 0, which can completely undermine the quality of the estimation. We have presented previously a preliminary graphic methodology to undertake this problem, and now some statistical procedures are suggested.

The designation of *statistical choice of extremal models* has been used to refer to the problem of choosing one of the three extremal types, and has been approached by countless authors under different conditions. It is usual to give preference in the null hypothesis to the simpler transitional Gumbel model, in the sense that the EVI is fixed as 0 as the frontier between the Fréchet and Weibull domains, between distributions with infinite and finite right endpoints.

Also, most common distributions belong to this domain, so it is reasonable to start from there. Regarding the alternative hypothesis, it can be in our interest to test a two-sided alternative

$$H_0 : \xi = 0 \qquad versus \qquad H_1^{(1)} : \xi \neq 0\,, \tag{3.14}$$

concluding simply if the distribution belongs or not to the Gumbel domain; or, if we aim to specify which of the other two max-domains is most adequate, we can test the one-sided alternatives

$$\begin{array}{llll} H_0 : \xi = 0 & versus & H_1^{(2)} : \xi < 0 & \text{for a Weibull alternative domain,} \\ H_0 : \xi = 0 & versus & H_1^{(3)} : \xi > 0 & \text{for a Fréchet alternative domain.} \end{array} \tag{3.15}$$

Let us then present some test procedures for these hypothesis, always considering the asymptotic level α. A more complete overview of tests for the statistical choice of extremal models can be found in Neves and Fraga Alves (2008), and details for each specific test statistic can be found in the references cited ahead.

Likelihood Ratio Test (LRT) This method for testing (3.14) is exposed with detail in Hosking (1984) and is based on the *Deviance* statistic. Being $(Y_1, \ldots, Y_m)$ a random sample of maxima i.i.d. to $Y \backsim G_\xi$ as defined in (2.14), let $l(\xi, \mu, \sigma)$ and $l(0, \mu, \sigma)$ denote resp. the unrestricted and the Gumbel-restricted log-likelihood functions with concern to the observed sample $(y_1, \ldots, y_m)$. The LRT statistic is the deviance

$$\mathbf{L} = -2\left(l(0, \hat{\mu}_{G_0}, \hat{\sigma}_{G_0}|Y_1, \ldots, Y_m) - l(\hat{\xi}_{G_\xi}, \hat{\mu}_{G_\xi}, \hat{\sigma}_{G_\xi}|Y_1, \ldots, Y_m)\right)$$

where $(\hat{\mu}_{G_0}, \hat{\sigma}_{G_0})$ and $(\hat{\xi}_{G_\xi}, \hat{\mu}_{G_\xi}, \hat{\sigma}_{G_\xi})$ are the ML estimates for the G_0 and G_ξ models, respectively.

Under H_0 and applying the Bartlett correction, as suggested in Hosking (1984), we have the modified statistic's approximation

$$\mathbf{L}^* = \frac{\mathbf{L}}{1 + \frac{2.8}{m}} \xrightarrow[m\to\infty]{d} Z \backsim \chi_1^2\,.$$

The null hypothesis is then rejected if $\mathbf{L}^* \geq \chi^2_{1,1-\alpha}$, where $\chi^2_{1,1-\alpha}$ denotes the χ_1^2 distribution's $(1-\alpha)$-quantile. The associated p-value can be calculated as $p(\mathbf{L}^*) = 1 - \chi_1^2(\mathbf{L}^*)$.

Rao's Score Test This test was first introduced by Tiago de Oliveira (1981) for testing both (3.14) and (3.15), and later furthered by Hosking (1984) for two-sided testing only. Considering $(\hat{\mu}_{G_0}, \hat{\sigma}_{G_0})$ the ML estimates for the G_0 model, the score statistic is obtained from $V(\xi|\mu, \sigma, y_1, \ldots, y_m)$ (the *score function* with respect to ξ) as follows:

$$V_m = \lim_{\xi\to 0} V(\xi|\hat{\mu}_{G_0}, \hat{\sigma}_{G_0}, Y_1, \ldots, Y_m) = \sum_{i=1}^{m}\left(\frac{1}{2}Z_i^2 - Z_i - \frac{1}{2}Z_i^2\,\exp(-Z_i)\right),$$

with $Z_i = \frac{Y_i - \hat{\mu}_{G_0}}{\hat{\sigma}_{G_0}}$ for $i = 1, \ldots, m$.

Under the null hypothesis, according to Tiago de Oliveira (1981) stands the approximation

$$V_m^* = \frac{V_m}{\sqrt{2.09797\, m}} \xrightarrow[m\to\infty]{d} Z \frown \mathcal{N}(0,1), \tag{3.16}$$

and equivalently, as presented by Hosking (1984),

$$V_m^{*2} = \frac{V_m^2}{2.09797\, m} \xrightarrow[m\to\infty]{d} Z^{(2)} \frown \chi_1^2\,. \tag{3.17}$$

The null hypothesis in the two-sided test (3.14) is then rejected if $|V_m^*| \geq z_{1-\frac{\alpha}{2}}$ or equivalently $V_m^{*2} \geq \chi^2_{1,1-\alpha}$, with associated p-values calculated as $p(V_m^*) = 2 - 2\,\Phi(|V_m^*|)$ and as $p(V_m^{*2}) = 1 - \chi_1^2(V_m^{*2})$. Note that $\Phi(.)$ denotes the standard Normal d.f. and $z_{1-\frac{\alpha}{2}}$ its $(1-\frac{\alpha}{2})$-quantile.

For the one-sided tests in (3.15), the rejection regions are given by $V_m^* \leq z_\alpha$ or $V_m^* \geq z_{1-\alpha}$ with associated p-values $p(V_m^*) = \Phi(V_m^*)$ or $p(V_m^*) = 1 - \Phi(V_m^*)$ resp. when dealing with the Weibull or Fréchet alternative domain.

Locally Asymptotically Normal (LAN) Test This test statistic can be applied to testing both (3.14) and (3.15) and has been thoroughly studied by Marohn (2000). Considering $(\hat{\mu}_{G_0}, \hat{\sigma}_{G_0})$ the ML estimates for the G_0 model, the LAN test statistic is given by

$$T_m = \frac{1}{3.451}\left(\frac{1.6449}{\sqrt{m}}\, S_{1,m} - \hat{\sigma}_{G_0}\,\frac{0.5066}{\sqrt{m}}\, S_{2,m} - \hat{\sigma}_{G_0}\,\frac{0.8916}{\sqrt{m}}\, S_{3,m}\right)$$

where $S_{1,m}$, $S_{2,m}$ and $S_{3,m}$ are the components of the *score function*, that is the first derivatives of the log-likelihood function of the GEV distribution with respect to each of the parameters, calculated at the point $\xi = 0$:

$$\begin{aligned}
S_{1,m} &= \sum_{i=1}^{m}\left(\frac{1}{2}Z_i^2 - Z_i - \frac{1}{2}Z_i^2\,\exp(-Z_i)\right),\\
S_{2,m} &= \sum_{i=1}^{m}\left(-\frac{1}{\hat{\sigma}_{G_0}} + \frac{1}{\hat{\sigma}_{G_0}} Z_i\,(1-\exp(-Z_i))\right),\\
S_{3,m} &= \sum_{i=1}^{m}\left(\frac{1}{\hat{\sigma}_{G_0}} - \frac{1}{\hat{\sigma}_{G_0}}\,\exp(-Z_i)\right),
\end{aligned}$$

with $Z_i = \frac{Y_i - \hat{\mu}_{G_0}}{\hat{\sigma}_{G_0}}$ for $i = 1, \ldots, m$.

Under H_0 we have the modified statistic's approximation

$$T_m^* = \frac{T_m}{0.6904} \xrightarrow[m\to\infty]{d} Z \frown \mathcal{N}(0,1)\,.$$

The null hypothesis in the two-sided test (3.14) is then rejected if $|T_m^*| \geq z_{1-\frac{\alpha}{2}}$, with associated p-value calculated as $p(T_m^*) = 2 - 2\,\Phi(|T_m^*|)$.

For the one-sided tests in (3.15), the rejection regions are given by $T_m^* \leq z_\alpha$ or $T_m^* \geq z_{1-\alpha}$ with associated p-values $p(T_m^*) = \Phi(T_m^*)$ or $p(T_m^*) = 1 - \Phi(T_m^*)$ resp. when dealing with the Weibull or Fréchet alternative domain.

Gumbel Statistic - as in Tiago de Oliveira and Gomes (1984) These authors presented a specific test for the hypothesis in (3.15) based on the *Gumbel Statistic*:

$$GS_m = \frac{Y_{m:m} - Y_{([m/2]+1):m}}{Y_{([m/2]+1):m} - Y_{1:m}}. \tag{3.18}$$

Under H_0 stands the approximation

$$GS_m^* = \frac{GS_m - \beta_m}{\alpha_m} \xrightarrow[m\to\infty]{d} W \backsim \Lambda$$

where the normalizing constants are

$$\alpha_m = \frac{1}{\log(\log(m))} \qquad \text{and} \qquad \beta_m = \frac{\log(m) + \log(\log(2))}{\log(\log(m)) - \log(\log(2))}.$$

The rejection regions are then given by $GS_m^* \leq \mathcal{G}_\alpha$ or $GS_m^* \geq \mathcal{G}_{1-\alpha}$ with associated p-values $p(GS_m^*) = \Lambda(GS_m^*)$ or $p(GS_m^*) = 1-\Lambda(GS_m^*)$ resp. when dealing with the Weibull or Fréchet alternative domain. Note that $\mathcal{G}_\varepsilon$ denotes the ε-quantile for the Gumbel distribution function $\Lambda(.)$.

Gumbel Statistic - as in Gomes and Fraga Alves (1996) In this paper it is suggested, for testing both (3.14) and (3.15), the application of the same *Gumbel Statistic* to the r top o.s.'s:

$$G_r = \frac{Y_{m:m} - Y_{(m-[\frac{r+1}{2}]+1):m}}{Y_{(m-[\frac{r+1}{2}]+1):m} - Y_{(m-r+1):m}}. \tag{3.19}$$

making the GS_m statistic in (3.18) a particular case of G_r when $r = m$.

The normalization constants suggested are also different from the ones considered in Tiago de Oliveira and Gomes (1984). Under H_0 stands, with the shown constants,

$$G_r^* = \frac{G_r - \beta_r}{\alpha_r} \xrightarrow[r\to\infty]{d} W \backsim \Lambda$$

$$\alpha_r = \frac{1}{\log(2)} \qquad \text{and} \qquad \beta_r = \frac{\log[\frac{r+1}{2}]}{\log(2)}.$$

The null hypothesis in the two-sided test (3.14) is then rejected if $G_r^* < \mathcal{G}_{\frac{\alpha}{2}}$ or $G_r^* > \mathcal{G}_{1-\frac{\alpha}{2}}$, with associated p-value calculated as $p(G_r^*) = 2\,\min\{\Lambda(G_r^*), 1 - \Lambda(G_r^*)\}$.

The rejection regions for (3.15) are given by $G_r^* \leq \mathcal{G}_\alpha$ or $G_r^* \geq \mathcal{G}_{1-\alpha}$ with associated p-values $p(G_r^*) = \Lambda(G_r^*)$ or $p(G_r^*) = 1-\Lambda(G_r^*)$ resp. when dealing with the Weibull or Fréchet alternative domain.

Another tool useful in approaching this problem is the goodness-of-fit tests, based on the empirical distribution function. Specifically in this context, the procedure is applied to testing the goodness-of-fit of the Gumbel model G_0 using the Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling test statistics, and was the focus of attention of Stephens (1976), (1977) and (1986).

Let it then be $(Y_1, \ldots, Y_m)$ the random sample of maxima from the underlying population X with d.f. F. The null hypothesis of this tests is

$$H_0 : \; F(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right)$$

meaning the Gumbel distribution with unknown location and scale parameters, making this a composite hypothesis. The alternative is not specified and thus this tests serve solely for analysing if a distribution from the Gumbel family would fit properly to the data, being equivalent to testing (3.14). The test statistics provide a sort of measure of the difference between the empirical d.f. and the best distribution in the null hypothesis, and as such the ML estimates of the parameters $(\hat{\mu}_{G_0}, \hat{\sigma}_{G_0})$ for the G_0 model are necessary.

The statistics are given as:

- Kolmogorov-Smirnov Statistic:

$$D_m = \max_{1 \le i \le m} \left\{ \left| G_0(Y_{i_m}|\hat{\mu}_{G_0}, \hat{\sigma}_{G_0}) - \frac{i}{m} \right|, \left| G_0(Y_{i_m}|\hat{\mu}_{G_0}, \hat{\sigma}_{G_0}) - \frac{i-1}{m} \right| \right\} \tag{3.20}$$

- Cramér-von Mises Statistic:

$$W_m^2 = \sum_{i=1}^{m} \left(G_0(Y_{i_m}|\hat{\mu}_{G_0}, \hat{\sigma}_{G_0}) - \frac{2i-1}{m} \right)^2 + \frac{1}{12\,m} \tag{3.21}$$

- Anderson-Darling Statistic:

$$\begin{aligned} A_m^2 = -m - \frac{1}{m} \sum_{i=1}^{m} \{ & (2i-1) \, \log\left(G_0(Y_{i_m}|\hat{\mu}_{G_0}, \hat{\sigma}_{G_0})\right) + \\ & + (2m+1-2i) \, \log\left(1 - G_0(Y_{i_m}|\hat{\mu}_{G_0}, \hat{\sigma}_{G_0})\right) \} \end{aligned} \tag{3.22}$$

The rejection of the null hypothesis happens for values of these test statistics that exceed a given quantile of asymptotic level α. The quantiles used are simulated and normally referred as the upper tail percentage points. The following tables represent a portion of the simulated quantiles that can be found in Chandra et al. (1981) for the Kolmogorov-Smirnov statistic and in Stephens (1977) for the Cramér-von Mises and Anderson-Darling statistics.

Table 3.1: Upper tail percentage points for the Kolmogorov-Smirnov statistic, modified for the Gumbel distribution.

		Upper tail significance level α			
Statistic	**m**	0.10	0.05	0.025	0.01
$\sqrt{m}\,D_m$	10	0.760	0.819	0.880	0.944
	20	0.779	0.843	0.907	0.973
	50	0.790	0.856	0.922	0.988
	∞	0.803	0.874	0.939	1.007

Chandra et al. (1981)

Table 3.2: Upper tail percentage points for the Cramér-von Mises and Anderson-Darling statistics, modified for the Gumbel distribution.

		Upper tail percentage points, α				
Statistic	**Modification**	0.75	0.90	0.955	0.975	0.99
W_m^2	$W_m^2(1+{}^{0.2}/\sqrt{m})$	0.073	0.102	0.124	0.146	0.175
A_m^2	$A_m^2(1+{}^{0.2}/\sqrt{m})$	0.474	0.637	0.757	0.877	1.038

Stephens (1977)

3.1.2 Exceedance Analysis – Peaks Over Threshold Method

This approach is more recent than the Gumbel methodology and is based on the theoretical framework described in Section 2.4. Instead of limiting our study to the single maximum value observed in each block, the POT methodology instructs us to choose an appropriately high threshold and, assuming enough observations fall beyond that level, fit a Generalized Pareto model (as defined in (2.33)) to those exceedances or to their transformation to the excesses. The Pickands-Balkema-de Haan Theorem (Theorem 2.4.1) shows us the duality between this approach and the classical method explored previously, since fitting a GEV model with EVI ξ to the sample of maxima is parallel to fitting a GP model with the same value of the shape parameter to the sample of excesses over a threshold. As such, we can obtain the same information on the tail of the distribution of the underlying population X through either method.

Consider the original sample $(X_1, \ldots, X_n)$ and let u be the chosen threshold. Let us denote N_u the number of observations $X_i > u$, that is, the number of excesses and the size of the sample we will now deal with. As in Section 2.4, we define the random variable of the conditional excesses over u as $Y|Y > 0$, with $Y := X - u$.

If in the BM method the choice of the block size was a delicate subject to keep in mind, the choice that has to be made attentively here is that of the threshold u. A procedure to assist this choice will be presented ahead.

As well as in the previous method, in the POT approach it is important that we position ourselves in the correct domain of attraction for the estimation, as illustrated in the previous section. And once again we can use preliminary graphical methodologies to get a sense of the value of the EVI, indicator of the most suitable domain. For example, an Exponentiality test (corresponding to the GP model when the EVI is null) can be performed by checking the presence (or lack) of linearity on an Exponential qq-plot: the trace between the empirical quantiles $y_{i:N_u}$ (sorted excesses) and the theoretical Exponential quantiles $-\log(1-p_i)$, with $p_i := {}^{i}/_{(N_u+1)}$ a possible definition of the plotting positions. The preliminary estimation of the scale parameter σ_u can be obtained from the qq-plot by the Least Squares method. More on this preliminary approach can be found in Gomes et al. (2013a).

The estimation techniques presented here will once again be the Maximum Likelihood and the Probability Weighted Moments methods, for reasons analogous to the ones stated in the BM case. Close to this, Bermudez and Kotz (2010a) and (2010b) constitutes a very complete overview of estimation techniques in the GP model.

3.1.2.1 Maximum Likelihood Estimation

Similarly to what was done for estimating the parameters of the GEV distribution, the ML procedure will be here used to find estimates for the parameters of the GP distribution. Under the conditions set above, we have the log-likelihood function with respect to $(y_1, ..., y_{N_u})$

$$\log L(\xi, \sigma_u | y_1, \ldots, y_{N_u}) = -N_u \log(\sigma_u) - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{N_u} \log\left(1 + \frac{\xi\, y_i}{\sigma_u}\right) \tag{3.23}$$

for values of the EVI $\xi \neq 0$ and $1 + \frac{\xi\, y_i}{\sigma_u} > 0$, $i = 1, \ldots, N_u$. For samples associated with a EVI $\xi = 0$, the log-likelihood expression is

$$\log L(0, \sigma_u | y_1, \ldots, y_{N_u}) = -N_u \log(\sigma_u) - \frac{1}{\sigma_u} \sum_{i=1}^{N_u} y_i \,. \tag{3.24}$$

Again, numerical maximization of this equations with respect to the parameters (ξ, σ_u) leads to the ML estimators $(\hat{\xi}, \hat{\sigma_u})$ for the GP family.

The consistency and asymptotic normality of these estimators have been established by Zhou (2009) and (2010) for values of the EVI $\xi > -1$.

3.1.2.2 Probability Weighted Moments Estimation

The other commonly used methodology is the PWM estimation, as referred before. Recall that the PWM estimators do not exist for values of the EVI $\xi \geq 1$ and that its implementation can yield inadmissable estimates, such as inference beyond a finite right endpoint. However, it has been shown that the method performs better than the ML method for small samples and specifically for EVI values $0 \leq \xi \leq 0.4$. In Hosking and Wallis (1987) it can be found the study of the estimators we now present, including proof of their asymptotically normal distributions.

Considering the probability weighted moments in their general definition in (3.3), being $(Y_1, \ldots, Y_{N_u})$ the random sample of i.i.d. variables from the GP family, the method uses the $M_{p,r,s}$ for $p = 1$, $r = 0$ and $s = 0, 1, 2, \cdots$, given by

$$M_{1,0,s} = \frac{\sigma_u}{(s+1)(s+1-\xi)}, \quad \xi < 1. \tag{3.25}$$

These moments can be estimated by

$$\hat{M}_{1,0,s} = \frac{1}{N_u} \sum_{i=1}^{N_u} \left(\prod_{k=1}^{s} \frac{(N_u - i - k + 1)}{(N_u - k)} \right) Y_{i:N_u} \,, \tag{3.26}$$

which allows us to obtain the PWM estimators for the shape and scale parameters:

$$\hat{\sigma}_u = \frac{2\,\hat{M}_{1,0,0}\hat{M}_{1,0,1}}{\hat{M}_{1,0,0} - 2\,\hat{M}_{1,0,1}} \qquad \text{and} \qquad \hat{\xi} = 2 - \frac{\hat{M}_{1,0,0}}{\hat{M}_{1,0,0} - 2\,\hat{M}_{1,0,1}} \,. \tag{3.27}$$

3.1.2.3 Estimation of Other Relevant Indicators

Having estimated the core parameters of the GP distribution as $(\hat{\xi}, \hat{\sigma_u})$ through the methods indicated previously, we can use these estimates to infer on the other interesting indicators already presented for the GEV distribution. Recall that the underlying population is $X \frown F$.

Exceedance Probability These probabilities are in this approach somewhat more complex to obtain, for they cannot be directly taken from the tail of the GP distribution with the estimated parameters – it must be weighted by the relative frequency of the excesses on the original sample (for more details on this matter, see Gomes et al. (2013a)):

$$\widehat{\overline{F(x)}} := \widehat{P(X > x)} = \frac{N_u}{n}\left(1 + \hat{\xi}\frac{x-u}{\hat{\sigma_u}}\right)^{-1/\hat{\xi}}; \tag{3.28}$$

Extremal Quantile With $U(.)$ the tail quantile function in 2.3.1, we have the theoretical extremal quantiles of exceedance probability p given by

$$U_{H_\xi}\left(\frac{1}{p}\right) = \begin{cases} \frac{\sigma_u}{\xi}\left(p^{-\xi} - 1\right), & \xi \neq 0 \\ -\sigma_u \log(p), & \xi = 0 \end{cases}. \tag{3.29}$$

For $\xi \neq 0$, these quantiles are estimated by:

$$\widehat{\chi_p} = \widehat{U\left(\frac{1}{p}\right)} = u + \frac{\hat{\sigma_u}}{\hat{\xi}}\left(\left(\frac{n\,p}{N_u}\right)^{-\hat{\xi}} - 1\right); \tag{3.30}$$

Right Endpoint In case the EVI is negative, $\xi < 0$, then we can estimate the value of the right endpoint as being the extremal quantile of probability 0:

$$\widehat{x^F} = \widehat{U(\infty)} = u + \frac{\hat{\sigma_u}}{\hat{\xi}}. \tag{3.31}$$

A more detailed study of these estimators can be found in Gomes et al. (2013a) or Vicente (2012).

3.1.2.4 Interval Estimation

Once again, the asymptotic normality of the ML or PWM estimators can be used to find confidence intervals for the parameters of the GP model. Nonetheless, as for the GEV distribution, more appealing interval estimates can be yield based on the profile log-likelihood function of the GP distribution with concern to ξ.

The $(1-\alpha) \times 100\%$ CI for the EVI produced is the same of the one presented for the GEV distribution in (3.13), with the proper profile log-likelihood function. See Beirlant et al. (2004) for more information on both types of interval estimation for the GP distribution.

3.1.2.5 Statistical Choice of Extreme Value Domains of Attraction

The choice of the domain of attraction to which the d.f. F belongs can also be done by testing procedures similar to the ones mention in Subsection 3.1.1.5, but now based on our excesses random variables, assumed $Y \backsim H_\xi$. Here, the transitional point of $\xi = 0$ corresponds to the Exponential model, and it will be tested against the same alternatives as before. As such, let us again consider the testing hypothesis in (3.14) and (3.15).

Since we are in the POT framework, the tests will be applied to the samples of the exceedances over the threshold u that as been used so far. We will denote it by $(W_1, \ldots, W_{N_u})$ i.i.d. to $W \backsim H_\xi$ whith H_ξ representing the reparameterization of (2.33) in order to $w = y + u$.

Let us then present some test procedures under these conditions, always considering the asymptotic level α. Again, in Neves and Fraga Alves (2008) it can be found a more complete overview of available tests of statistical choice in the POT framework, and details for each specific test statistic can be found in the references cited ahead.

Likelihood Ratio Test This test is analogous to the one presented in the GEV context, and can be applied to testing the two-sided hypothesis (3.14). Let $l(\xi, u, \sigma_u)$ and $l(0, u, \sigma_u)$ denote resp. the unrestricted and the Exponential-restricted log-likelihood functions with concern to the observed sample $(w_1, \ldots, w_{N_u})$. The LRT statistic is the deviance

$$\mathbf{L} = -2\left(l(0, u, \hat{\sigma}_{u,H_0}|W_1, \ldots, W_{N_u}) - l(\hat{\xi}_{H_\xi}, u, \hat{\sigma}_{u,H_\xi}|W_1, \ldots, W_{N_u})\right)$$

where $\hat{\sigma}_{u,H_0}$ and $(\hat{\xi}_{H_\xi}, \hat{\sigma}_{u,H_\xi})$ are the ML estimates for the H_0 and H_ξ models, respectively. Under the H_0 hypothesis and applying the Bartlett correction as suggested in Reiss and Thomas (2007) we have the modified statistic's approximation

$$\mathbf{L}^* = \frac{\mathbf{L}}{1 + \frac{4}{N_u}} \xrightarrow[N_u \to \infty]{d} Z \backsim \chi^2_1 \,.$$

The null hypothesis is then rejected if $\mathbf{L}^* \geq \chi^2_{1,1-\alpha}$, where $\chi^2_{1,1-\alpha}$ denotes the χ^2_1 distribution's $(1-\alpha)$-quantile. The associated p-value can be calculated as $p(\mathbf{L}^*) = 1 - \chi^2_1(\mathbf{L}^*)$.

G_{N_u} Test Statistic This statistic for testing the one-sided hypothesis in (3.15) was proposed and discussed in Gomes and van Monfort (1986) and is constructed from the sample maximum and median as

$$G_{N_u} = \frac{W_{N_u:N_u}}{W_{\left(\frac{[N_u]}{2}+1\right):N_u}}.$$

Under H_0 stands the approximation

$$G^*_{N_u} = \log(2)\, G_{N_u} - \log(N_u) \xrightarrow[N_u \to \infty]{d} Z \backsim \Lambda$$

and as such the rejection regions are then given by $G^*_{N_u} \leq \mathcal{G}_\alpha$ or $G^*_{N_u} \geq \mathcal{G}_{1-\alpha}$ with associated p-values $p(G^*_{N_u}) = \Lambda(G^*_{N_u})$ or $p(G^*_{N_u}) = 1 - \Lambda(G^*_{N_u})$ resp. when dealing with the Weibull or Fréchet alternative domain.

The authors Gomes and van Monfort (1986) also presented simulated quantiles for this statistic for dealing with small samples, but since these will not be applied in our case study (given the considerable size of our sample) their exposition was dismissed from this dissertation.

T_{N_u} **Test Statistic** This statistic was presented in the earlier mention paper from Marohn (2000) and can be used in testing either (3.14) or (3.15) hypothesis. Consider $S_W^2 = \frac{1}{N_u}\sum_{i=1}^{N_u}(W_i - \overline{W})^2$ the variance of the exceedances sample. The test statistic is defined as

$$T_{N_u} = \frac{1}{2}\left(\frac{S_W^2}{(\overline{W}-u)^2} - 1\right).$$

Under H_0, i.e., the validity of the Exponential model, we have the modified statistic's approximation

$$T_{N_u}^* = \sqrt{N_u}\, T_{N_u} \underset{N_u\to\infty}{\xrightarrow{d}} Z \frown \mathcal{N}(0,1)\,.$$

The null hypothesis in the two-sided test (3.14) is then rejected if $|T_{N_u}^*| \geq z_{1-\frac{\alpha}{2}}$, with associated p-value calculated as $p(T_{N_u}^*) = 2 - 2\,\Phi(|T_{N_u}^*|)$.

For the one-sided tests in (3.15), the rejection regions are given by $T_{N_u}^* \leq z_\alpha$ or $T_{N_u}^* \geq z_{1-\alpha}$ with associated p-values $p(T_{N_u}^*) = \Phi(T_{N_u}^*)$ or $p(T_{N_u}^*) = 1 - \Phi(T_{N_u}^*)$ resp. when dealing with the Weibull or Fréchet alternative domain.

The author showed that for the two-sided test in (3.14) the test statistic is biased, an has very poor power for samples up to moderate sizes ($N_u < 500$), only providing reasonable results when working with larger sample sizes.

Similarly to what was shown for the GEV, also for testing the GP goodness-of-fit tests can be applied, that work in the same way as described before. The difference lays in the hypothesis to be tested, since the test statistics used are the same. Therefore, the Kolmogorov-Smirnov test statistic will be used for testing the fit of an Exponential distribution to the data, while the Cramér-von Mises and Anderson-Darling test statistics will be used in testing the fit of the Generalized Pareto distribution. In both cases we have a composite null hypothesis, since the parameters are unknown, and thus we will need to know their ML estimates. The rejection of the null hypothesis happens for values of these statistics that exceed a given quantile of asymptotic level α. Tables referring to the simulated quantiles for each statistic will be presented.

Since we have the GP distribution (an its Exponential particular case for $\xi = 0$) parameterized in order of the excess variable Y in (2.33), the tests will be described for the sample $(Y_1, \ldots, Y_{N_u})$. The usage of the Kolmogorov-Smirnov test in an Exponential context was studied by Lilliefors (1969), and the application of the Cramér-von Mises and Anderson-Darling test statistics to a GP null hypothesis is described in Choulakian and Stephens (2001). According to these authors and for these hypothesis, the statistics are given by:

- Kolmogorov-Smirnov Statistic:

$$D_{N_u} = \max_{1\leq i\leq N_u}\left\{\left|1 - \exp\left(-\frac{Y_{i:N_u}}{\hat{\sigma_u}}\right) - \frac{i}{N_u}\right|, \left|1 - \exp\left(-\frac{Y_{i:N_u}}{\hat{\sigma_u}}\right) - \frac{i-1}{N_u}\right|\right\} \tag{3.32}$$

with $\hat{\sigma_u}$ the ML estimate of the Exponential's scale parameter σ_u; simulated critical values for D_{N_u} can be found in Table 3.3;

- Cramér-von Mises Statistic:

$$W^2_{N_u} = \sum_{i=1}^{N_u} \left(H_{\hat{\xi}}(Y_{i:N_u}|\hat{\sigma}_{u,H_\xi}) - \frac{2i-1}{N_u} \right)^2 + \frac{1}{12\,N_u} \tag{3.33}$$

- Anderson-Darling Statistic:

$$\begin{aligned} A^2_{N_u} = -N_u - \frac{1}{N_u} \sum_{i=1}^{N_u} \{ (2i-1)\, \log\left(H_{\hat{\xi}}(Y_{i:N_u}|\hat{\sigma}_{u,H_\xi}) \right) + \\ + (2N_u + 1 - 2i)\, \log\left(1 - H_{\hat{\xi}}(Y_{i:N_u}|\hat{\sigma}_{u,H_\xi}) \right) \} \end{aligned} \tag{3.34}$$

with $H_\xi(.)$ the GP distribution defined in (2.33) and $(\hat{\xi}, \hat{\sigma}_{u,H_\xi})$ the ML estimates for its shape and scale parameters (ξ, σ_u); simulated critical values for $W^2_{N_u}$ and $A^2_{N_u}$ can be found in Tables 3.4 and 3.5 respectively.

Table 3.3: Simulated critical values of the Kolmogorov-Smirnov statistic adapted to the Exponential distribution with unknown parameters.

		Level of Significance for D_{N_u}		
Statistic	N_u	0.10	0.05	0.01
D_{N_u}	5	0.406	0.442	0.504
	10	0.295	0.325	0.380
	15	0.244	0.269	0.315
	20	0.212	0.234	0.278
	30	0.174	0.192	0.226
	> 30	$\frac{0.96}{\sqrt{N_u}}$	$\frac{1.06}{\sqrt{N_u}}$	$\frac{1.25}{\sqrt{N_u}}$

Lilliefors (1969)

Table 3.4: Simulated critical values of the Cramér-von Mises statistic adapted to the GPd with unknown parameters.

		Upper-Tail Asymptotic Percentage Points		
Statistic	ξ	0.10	0.05	0.01
$W^2_{N_u}$	0.9	0.094	0.115	0.165
	0.5	0.101	0.124	0.179
	0.1	0.116	0.144	0.210
	0	0.124	0.153	0.224
	-0.1	0.129	0.160	0.236
	-0.5	0.174	0.222	0.338

Choulakian and Stephens (2001)

Table 3.5: Simulated critical values of the Anderson-Darling statistic adapted to the GPd with unknown parameters.

		Upper-Tail Asymptotic Percentage Points		
Statistic	ξ	0.10	0.05	0.01
$A^2_{N_u}$	0.9	0.641	0.771	1.086
	0.5	0.685	0.830	1.180
	0.1	0.766	0.935	1.348
	0	0.796	0.974	1.409
	-0.1	0.831	1.020	1.481
	-0.5	1.061	1.321	1.958

Choulakian and Stephens (2001)

3.1.2.6 Choice of Threshold

As stated before in this dissertation, a very important question in this approach is how high to take the threshold level u. The problem is in finding a level that balances the big variance of the estimators that occurs for too large values of u and their significant bias that occurs for much smaller values of this threshold. Although this problem has been studied in the specialized literature, under heuristic, resampling and theoretical approaches, it remains still controversial and with no satisfactory global solution. We will present and latter apply a pragmatic methodology proposed by Davison and Smith (1990) based on the study of the *mean excess function* (m.e.f.), defined as

$$e(u) := E[X - u | X > u], \quad \text{if} \quad E[X] < \infty\,, \tag{3.35}$$

and its empirical counterpart, based on the originally observed sample $(x_1, \ldots, x_n)$ as

$$\hat{e}_n(u) := \frac{\sum_{i=1}^n x_i\, \mathbb{I}_{(u,\infty)}(x_i)}{\sum_{i=1}^n \mathbb{I}_{(u,\infty)}(x_i)} - u, \qquad \text{with} \qquad \mathbb{I}_{(u,\infty)} = \begin{cases} 1, \text{ if } x_i \in (u,\infty) \\ 0, \text{ if } x_i \in (-\infty, u] \end{cases}. \tag{3.36}$$

If the GP assumption is correct, then the m.e.f. takes the form

$$e(u) := E[X - u | X > u] = E[Y | Y > 0] = \frac{\sigma_u + \xi\, u}{1 - \xi}, \quad \text{if} \quad \xi < 1 \tag{3.37}$$

and the plot of $\hat{e}_n(u)$ against u should follow a straight line with intercept $\frac{\sigma_u}{1-\xi}$ and slope $\frac{\xi}{1-\xi}$, suggesting both graphical estimates of the parameters and a fit test based on the linearity of the plot. The sample mean excess plot is usually constructed by considering as possible threshold values the empirical quantiles from the sample of X, plotting the curve of $(X_{n-k:n}, \hat{e}_n(X_{n-k:n}))$ for $k = 1, \ldots, n-1$, and allowing for the empirical m.e.f. in (3.36) to be rewritten as

$$\hat{e}_n(x_{n-k:n}) := \frac{1}{k} \sum_{j=1}^{k} x_{n-j+1:n} - x_{n-k:n}\,.$$

Ideally, if the data truly comes from a GP distribution, then the plot of the empirical m.e.f. would be completely linear. But even when it is the case, linearity is rarely absolute, specially towards the highest levels of the threshold, where there is a very small number of excesses being averaged, making it common practice when constructing this plot to omit the last few points so they don't disrupt the plot. So, this technique consists in finding the point on the plot such that a linear pattern is visible to its left, corresponding to the desired threshold u.

This problem of finding the threshold u is paralleled by another – the choice of the number k of top o.s.'s to utilize in the semi-parametric estimation, and specifically the random threshold $X_{n-k:n}$ above which the ordered excesses are calculated. This methodology, detailed ahead, has been known since firstly named in Araújo Santos et al. (2006) as the *Peaks Over Random Threshold* (PORT) method, as the threshold considered is given by a random o.s..

Other methods for choosing the level of exceedance can be found in Coles (2001) or Gong (2012).

3.1.3 Multidimensional Approach – Largest Yearly Observations Modelling

Another way of dealing with the intrinsic difficult to EVA related to the limited amount of data for model estimation, alternative to the POT methodology, is the Largest Observations per block approach. As in the Gumbel method, we need first to consider our collection of data to be divided in m blocks. Here stands the issue of block size choice, the trade between variance and bias of the resulting estimators, usually resolved by the pragmatic choice of considering blocks corresponding to yearly spans of time.

The method then consists in modeling the largest k order statistics of each block with the limiting behaviour described in Theorem 2.5.2, known as Multivariate GEV model or Extremal Process GEV. This allows us to increase the amount of information used in the estimation without having to increase the number of observations. Note that, as stated, considering $k = 1$ corresponds to the Gumbel or BM methodology. There is also the possibility of considering different block sizes for each block, which is common when, for instance, a certain year has less recorded values than the others, for some reason, but we will here assume the block sample size is the same for all blocks, k. This model is also known as the Multidimensional GEV model.

For each block $j = 1, \ldots, m$, there is then a sample of the form $\underset{\sim}{x}_j = (x_{1,j}, \ldots, x_{k,j})$ with the order $x_{1,j} > \ldots > x_{k,j}$ such that $x_{1,j}$ corresponds to the j^{th} block maximum. Then, the log-likelihood function for this model (absorbing the unknown scaling coefficients into location and scale parameters) is given as

$$\begin{aligned} \log L(\xi, \mu, \sigma | \underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_m) = \sum_{j=1}^{m} \Bigg\{ & - \left(1 + \xi \frac{x_{k,j} - \mu}{\sigma}\right)^{-1/\xi} - \\ & - \sum_{i=1}^{k} \left(\log(\sigma) + \left(\frac{1}{\xi} + 1\right) \log\left(1 + \xi \frac{x_{k,j} - \mu}{\sigma}\right)\right) \Bigg\} \end{aligned} \tag{3.38}$$

for values of the tail index $\xi \neq 0$ and $1 + \xi \frac{x_{i,j}-\mu}{\sigma} > 0$ for $i = 1, \ldots, k$ and $j = 1, \ldots, m$, and as

$$\log L(0, \mu, \sigma | \underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_m) = \sum_{j=1}^{m} \left\{ -\exp\left(-\frac{x_{k,j}-\mu}{\sigma}\right) - \sum_{i=1}^{k} \left(\log(\sigma) + \frac{x_{k,j}-\mu}{\sigma}\right) \right\} \tag{3.39}$$

for the case $\xi = 0$.

The log-likelihood functions in (3.38) and (3.39) can be numerically maximized in order to obtain the ML estimates for the EVI ξ, the location μ and scale σ. These correspond to the same parameters of the GEV distribution for BM, but with the incorporation of extra information of the observed extreme data. Thus, their interpretation is changeless, but the precision of the estimates is expected to be greater, due to the inclusion of more information. These estimates can then be used in the same fashion as presented for the BM approach in estimating other indicators of extreme values, such as exceedance probabilities and the right endpoint for the case $\xi < 0$. Standard asymptotic likelihood theory also allows for the construction of approximate confidence intervals.

In Gomes (1981), paper where this approach was firstly introduced, are presented explicit expressions for the ML estimators for the location and scale parameters under each of the extremal types models, Gumbel, Fréchet and Max-Weibull. Related to this, Smith (1986) also considered the same type of multidimensional approach. Practical applications of this theory can be found in Coles (2001) and Fawcett (2012).

3.1.4 About Non-Stationarity

All the theory and methodologies presented thus far are based in the main assumption that the observed values can be acceptably considered as realizations of an i.i.d. sample $(X_1, \ldots, X_n)$, implying the assumption of stationarity of the underlying process, leading us to constant parameter estimators for the distributions of extremes addressed – the GEV in the Gumbel approach, the GP in the POT approach, the GEV extremal process in the LO approach.

However, in many common applications, such assumption is unrealistic, as the process does not remain the same as time changes, possibly presenting trend, seasonality or even volatility. It is reasonable to presume that the temporal dynamics of the entire series and that of its extremes are closely tied, challenging the suitability of the modeling techniques used when a constant distribution through time is assumed. In practice, it is common to undertake some pragmatic deviations based on the type of non-stationarity observed, for no general theory can be established for these kinds of processes. Some specific results exist, but these are typically too restrictive to describe non-stationary patterns in real series. As such, the common approach is to use the standard extreme value models as basic templates to work upon, modifying them in order to incorporate those new features.

Many studies have been devoted to analysing the occurrence of temporal trends and cycles, using various techniques. To refer only a few, see, for example, the works in the environmental field by Smith (1989), using point process characterizations, Renard et al. (2006) with a Bayesian

approach, Méndez et al. (2006), Nogaj et al. (2007), Beguería et al. (2011) or most recently Fraga Alves (2015) and Vanem (2015). In de Haan et al. (2015) tail trend detection is investigated in the context of heavy rainfall, and in the field of sports, non-stationarity was considered by Stephenson and Twan (2013). Einmahl et al. (2016) is a more recent work on the inference for non-stationary models, with an application in financial series.

Note that non-stationarity is here understood as variation related to the intuitive variable *time*, but for different types of series could be taken as related to any appropriate covariate or index, such as the greenhouse gas emissions for environmental data. As a matter of fact it is appropriate to suggest that non-stationarity is more common than stationarity in environmental series such those of climate extremes, and the recommendation made by Nogaj et al. (2007) is to make use of the methodology hereafter presented unless the assumption of stationarity can be proven for each data series.

As stated, when the observed non-stationarity is in the form of trend and/or seasonality, the extreme value models are still useful, being that time dependant parameters can be the answer to specific problems. That is, we consider that a functional relationship exists between the model parameters and time, like a linear or log-linear relation which are preferable for trend modeling. Consider then the models (2.14) (with the family parametrization in (2.18)), (2.33) and (2.36) previously presented.

Modeling non-stationarity in the GEV model can be as simple as taking the parameters as time dependant $(\xi(t), \mu(t), \sigma(t))$. For example, a steady change in the sample of maxima can me modeled by a linear trend in the location parameter as $\mu(t) = \beta_0 + \beta_1 t$ with β_0 and β_1 real constant parameters to be estimated. Estimation for this modified $G_{\xi(t)}(.|\mu(t), \sigma(t))$ model can be made analogously to the stationary case through the ML method, where the log-likelihood function incorporates the functional parameters:

$$\log L(\xi(t), \mu(t), \sigma(t)|y_1, \ldots, y_m) = -m \log(\sigma(t)) - \left(\frac{1}{\xi(t)} + 1\right) \sum_{i=1}^{m} \log\left(1 + \xi(t)\frac{y_i - \mu(t)}{\sigma(t)}\right) - \sum_{i=1}^{m} \left(1 + \xi(t)\frac{y_i - \mu(t)}{\sigma(t)}\right)^{-\frac{1}{\xi(t)}} \tag{3.40}$$

for values of the EVI $\xi(t) \neq 0$ and $1 + \xi(t)\frac{y_i - \mu(t)}{\sigma(t)} > 0$, for $i = 1, \ldots, m$ and for every t. For samples associated with a EVI $\xi(t) = 0$, the log-likelihood expression is

$$\log L(0, \mu(t), \sigma(t)|y_1, \ldots, y_m) = -m \log(\sigma(t)) - \sum_{i=1}^{m} \exp\left(-\frac{y_i - \mu(t)}{\sigma(t)}\right) - \sum_{i=1}^{m} \frac{y_i - \mu(t)}{\sigma(t)}. \tag{3.41}$$

The ML estimates are then numerically obtained by maximization of these log-likelihood functions, as usual.

When considering the non-stationary POT methodology and the GP model, in Nogaj et al. (2007) two approaches are suggested: the nonparametric and the parametric non-stationary POT models. Note that the terms nonparametric and parametric refer only to how time dependence of the model parameters was modelled, but in both cases the GP model is assumed as the

underlying behaviour.

The nonparametric non-stationary POT model consists in dividing the sample in several periods, with different behaviours. This is usually useful in modeling the presence of seasonality, the one inherent to climatic processes such as rainfall. The model is then given by

$$(X - u|X > u) \backsim GP(\xi_{s(t)}, \sigma_{u,s(t)}) \tag{3.42}$$

where $(\xi_{s(t)}, \sigma_{u,s(t)})$ are the GP parameters for the time period $s(t)$, which can be viewed as the season in which time t falls. These parameters are assumed to be independent from the ones for other seasons, so they can be directly estimated from the subsample of data for that period. It might also be indicated to consider different thresholds $u_{s(t)}$ for each season. The issue with this method is the suitable segregation into time periods or seasons of the original sample.

The parametric non-stationary POT model is derived from the stationary POT methodology in an analogous way to how $G_{\xi(t)}(.|\mu(t), \sigma(t))$ was derived from the stationary GEV model – using functional relations between time and the model parameters $(\xi(t), \sigma_u(t))$. One useful relation commonly considered for modeling trend through time in a non-stationary POT framework is $\sigma_u(t) = \exp(\beta_0 + \beta_1 t)$ with β_0 and β_1 real constant parameters to be estimated, the exponential function here being useful to ensure the positivity of the scale parameter.

As no data sub-sampling is required with this parametrization, all the exceedances or excesses are used in fitting the model, this being the main advantage of the parametric over the nonparametric model, and ML estimates of the parameters can be easily obtained from the modified log-likelihood functions

$$\log\, L(\xi(t), \sigma_u(t)|y_1, \ldots, y_{N_u}) = -N_u\, \log(\sigma_u(t)) - \left(\frac{1}{\xi(t)} + 1\right) \sum_{i=1}^{N_u} \log\left(1 + \frac{\xi(t)\, y_i}{\sigma_u(t)}\right) \tag{3.43}$$

for values of the EVI $\xi(t) \neq 0$ and $1 + \frac{\xi(t)\, y_i}{\sigma_u(t)} > 0$, for $i = 1, \ldots, m$ and for every t, and for samples associated with a EVI $\xi(t) = 0$

$$\log\, L(0, \sigma_u(t)|y_1, \ldots, y_{N_u}) = -N_u\, \log(\sigma_u(t)) - \frac{1}{\sigma_u(t)} \sum_{i=1}^{N_u} y_i\,. \tag{3.44}$$

The case of the non-stationary Largest Observations method is dealt with in a completely similar manner to what has been shown for the BM and POT methods. One simply takes the GEV extremal process with time dependant parameters $(\xi(t), \mu(t), \sigma(t))$, and the consequential direct modifications to the likelihood function used in the ML estimation of such parameters (by intermediate of the estimation of the real constants that determine the functional relations in the parameters).

It is usually very hard to model the shape parameter ξ as a function of time, so the hypothesis of time invariance of the EVI is a reasonable one. Regular trends of this parameter, such as linear trends, have minor statistical significance and are generally refuted by goodness-of-fit tests, as shown in Nogaj et al. (2006) for the POT approach. As such, time dependance of the tail index

is rarely considered.

Since for each modelling framework there exist countless combinations of the model parameters as functions of time (or other appropriate covariate), a procedure for choosing the *best* model is required. Here, the *best* model is the simplest, most parsimonious one which explains the largest amount of data variability as possible. If a linear trend in the location parameter sufficiently explains the variation in the maxima sample, for example, perhaps considering a quadratic trend will not add any significant information to the model. But perhaps it will. In this case, the class of quadratic trends includes the class of linear ones, hence both models would be nested.

The Likelihood Ratio test, based in the Deviance statistic, is a useful tool in choosing models in these conditions. Consider one of the approaches above (non-stationary BM, POT or LO) and two nested fitted models $\mathcal{M}_0$ and $\mathcal{M}_1$, where $\mathcal{M}_0$ is the simpler more restrictive model – $\mathcal{M}_0 \subset \mathcal{M}_1$. Being $l_0(\mathcal{M}_0)$ and $l_1(\mathcal{M}_1)$ the maximized log-likelihood functions of said models, the values of the deviance statistic $\mathbf{L} = -2\{l_0(\mathcal{M}_0) - l_1(\mathcal{M}_1)\}$ will allow us to decide if the more complex model significantly improves the description of the data, if $\mathbf{L}$ is fairly large. The approximation of $\mathbf{L}$ to a χ^2_q distribution, where q is the difference in the dimensionality (number of parameters) of the two models, dictates the rejection at the significance level α of the null hypothesis – both models are statistically equivalent $H_0 : \mathcal{M}_0 = \mathcal{M}_1$ – if $\mathbf{L}_{obs} > \chi^2_{q,1-\alpha}$, where $\chi^2_{q,1-\alpha}$ denotes the χ^2_q distribution's $(1-\alpha)$-quantile.

Having chosen the appropriate model and estimated its functional parameters, the inference on the other extreme values indicators must be carefull, since all the estimates will be indexed in time. The estimation makes use of the corresponding expressions provided in the previous sections. Thus, any estimation can only be made based on a temporal horizon. It is wise to limit the inference to short term predictions, since the presence of non-stationarity itself dictates the changing nature of the series.

Think, for example, of an estimated positive trend in sea water level. A thoughtless long term prediction based on this trend can lead us to believe that in a relatively close future, all planet will be submerged, and that in a relatively close past, no water existed at all, since the maximum sea level is estimated to be rising at a given rate **now**. Therefore, careful attention must be taken for such estimations.

Further information on this topic, including model diagnostics, is detailed in Coles (2001) and Fawcett (2012).

3.2 Semi-Parametric Approach

In this section we will detail an alternative statistical methodology to the parametric inference presented before, whose central concept was the existence of a class of parametric models of extremes that appropriately described the r.v. X at hand. Regarding this type of approach some questions have arisen concerning the extend of the validity of its assumptions – binding the data to an asymptotic parametric model that may be too rigid or idealistic. There is also the question of purpose in the analysis, since in most applications of EVT the goal is to describe the behaviour of extreme values, the frequency and/or magnitude of rare events, not to much modeling the data at the expense of a strict theoretical distribution.

The study of the semi-parametric approach that we from here on out present was lead firstly by the works of Hill (1975) and Pickands (1975), and the development of the regular variation theory laid down by de Haan (among others) is now the theoretical setup under which most advances in this methodology are made – usefully condensed in de Haan and Ferreira (2006). The designation emphasises that the focus is just the same on the estimation of the extreme value parameters and indicators, just here under only partial assumptions on the underlying distribution function F of X. Unlike under the parametric framework, in semi-parametric inference there is no model being fitted to the data, only some imposed conditions and restrictions about the tail behaviour of F, upon which we wish to infer, and for that reason the alternative denomination of "parametric on the tail" methodology is suggested in Fraga Alves (1999).

As such, rather than fitting an extreme value model to the whole sample, the assumption made on F is that the *First Order Condition*, defined in the second statement of Theorem 2.3.1 in section 2.3, is satisfied, i.e., $F \in \mathcal{D}_{\mathcal{M}}(G_\xi)$ for some real value of the EVI ξ. Hence, the EVI continues to be the main parameter of the EVT to be estimated, regardless all focus being on the domain of attraction rather than on the limiting distribution. Even though the validity of the First Order Condition is enough for developing the semi-parametric theory, some properties of the estimators it yields require the validity of the *Second Order Condition*, defined in Definition 2.3.3, in section 2.3, such as the property of asymptotic normality, as will be shown ahead.

In this setup, all the inference can be based on the top $k+1$ values above a random threshold from the original sample of size n. We denote the referred decreasing ordered sample by $(X_{(1)}, \ldots, X_{(k+1)})$, where $X_{(i)}$ represents the i^{th} largest order statistic $X_{n-i+1:n}$, $i = 1, \ldots, k+1$. Formally, the determination of such threshold is dictated by k an *intermediate sequence* of positive integers such that

$$k \equiv k_n \to \infty \qquad \text{with} \qquad \frac{k_n}{n} \to 0 \quad \text{as} \quad n \to \infty\,, \tag{3.45}$$

which means the random threshold $X_{n-k:n}$ corresponds to an *intermediate o.s.* in the sense of the definition given in section 2.1. Some restrictions are imposed on the rate of increase of k and n, according to the Second Order Condition. The speed with which $k_n \to \infty$ must be in accordance with the rate of convergence of the First Order Condition, and as such quantified by the A function in (2.23), for what the choice of the sequence k_n is controlled by this condition.

One can easily realize the importance of an adequate choice of k, but the way of determining it is not so simple, or even consensual. It is a problem parallel to the choice of threshold u in the POT methodology. Generally, choosing too small values of k leads to high variability of the estimators, while too large values of k leads to a large bias.

Recall that, by *statement* 4 of Theorem 2.3.1 exists a positive function f, for example $f(t) = a\left(\frac{1}{\overline{F}(t)}\right)$, such that

$$\lim_{t\uparrow x^F} \frac{\overline{F}\left(t + xf(t)\right)}{\overline{F}(t)} = (1+\xi x)^{-1/\xi}, \quad \forall x:\ 1+\xi x > 0 \tag{3.46}$$

and this is equivalent to the First Order Condition. Considering the GP distribution in (2.33) and the approximation (2.34) given by the Pickands-Balkema-de Haan Theorem, we can equivalently write

$$\lim_{t\uparrow x^F} P\left[\frac{X-t}{f(t)} > x | X > t\right] = (1+\xi x)^{-1/\xi} = \overline{H}_\xi(x)\,.$$

Hence, from a high enough level t is valid

$$P\left[X > t + f(t)x\right] \approx P\left[X > t\right]\overline{H}_\xi(x)$$

or even

$$P\left[X > x\right] = P\left[X > t + f(t)\frac{x-t}{f(t)}\right] \approx P\left[X > t\right]\overline{H}_\xi\left(\frac{x-t}{f(t)}\right), \quad x > t,$$

i.e.

$$\overline{F}(x) \approx \overline{F}(t)\left\{1 - H_\xi\left(\frac{x-t}{f(t)}\right)\right\}, \quad x > t. \tag{3.47}$$

Considering then the intermediate o.s. threshold $t = X_{(k+1)}$, the empirical distribution function defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(-\infty,x]}(x_i) \tag{3.48}$$

where $\mathbb{I}(.)$ is the Indicator function defined in (3.36), and the approximation $F(x) \approx F_n(x)$ we have

$$f(X_{(k+1)}) = a\left(\frac{1}{1-F(X_{(k+1)})}\right) \approx a\left(\frac{n}{k}\right). \tag{3.49}$$

Thus, from (3.47) with $t = X_{(k+1)}$ and having (3.49), we can write the approximation for the tail of F above the random threshold as

$$\overline{F}(x) \approx \frac{k}{n}\left\{1 - H_\xi\left(\frac{x - X_{(k+1)}}{a\left(\frac{n}{k}\right)}\right)\right\}, \quad x > X_{(k+1)}. \tag{3.50}$$

This approximation is the base for the semi-parametric inference to be performed on the tail of F, and as such it is crucial to estimate the EVI ξ but also the normalizing scaling constant given by the value of the function $a(.)$ at $\frac{n}{k}$. The location normalizing coefficient here is given by the random threshold $X_{(k+1)}$.

Further detailing on this results are described in de Haan and Ferreira (2006), Vicente (2012) and Rosário (2013).

3.2.1 Statistical Tests for the Extreme Value Index Sign

As stated above, the results comprising the semi-parametric approach are set under the assumption of validity of the First Order Condition, meaning the validity of $F \in \mathcal{D}_{\mathcal{M}}(G_\xi)$ for some real value of ξ. Therefore, the EVI is again the highlighted parameter as it determines the type of decay of the tail of F. Having this information can help us select the appropriate estimation procedure to be applied, considering some of the estimators to be presented ahead can be simplified or even discarded by previous knowledge of the EVI's sign. Subsequently, it is advisable to statistically test for the EVI's sign to assure that the estimation performed will not be meaningless.

Among the vast literature regarding semi-parametric methodologies, many proposals for testing procedures targeted at the selection of the correct max-domain of attraction exit. In Hüsler and Peng (2008) and Neves and Fraga Alves (2008) the authors give a general overview of some of the most well-known tests in this context. We will in this section follow some existing testing procedures for pre-testing the sign of ξ that can be applied regardless of the estimation of the EVI in itself, meaning that any previous estimation is not necessary.

Similarly to what happens in the parametric context, it is usual to give preference in the null hypothesis to the transitional case that corresponds to the Gumbel max-domain of attraction, in the sense that it constitutes the frontier between the Fréchet and Weibull domains, between distributions with lighter tails and finite right endpoint and distributions with heavier tails and infinite right endpoint. As said before, most common distributions belong to this domain and as such the tests for choosing the most appropriate domain of attraction for the tail distribution in this setup are used on the hypothesis

$$H_0 : \ F \in \mathcal{D}_{\mathcal{M}}(G_0) \qquad versus \qquad H_1^{(1)} : \ F \in \mathcal{D}_{\mathcal{M}}(G_\xi)_{\xi \neq 0} \tag{3.51}$$

or rather on the one-sided alternatives

$$\begin{aligned} &H_0 : \ F \in \mathcal{D}_{\mathcal{M}}(G_0) \qquad versus \qquad H_1^{(2)} : \ F \in \mathcal{D}_{\mathcal{M}}(G_\xi)_{\xi<0} \qquad \text{for a Weibull alternative domain,} \\ &H_0 : \ F \in \mathcal{D}_{\mathcal{M}}(G_0) \qquad versus \qquad H_1^{(3)} : \ F \in \mathcal{D}_{\mathcal{M}}(G_\xi)_{\xi>0} \qquad \text{for a Fréchet alternative domain.} \end{aligned} \tag{3.52}$$

The three testing procedures hereby presented depend on the observations from the sample that fall beyond a random threshold $X_{(k+1)} := X_{n-k:n}$, with test statistics that are based on the k excesses over that level, namely

$$Z_i := X_{(i)} - X_{(k+1)}, \quad i = 1, \ldots, k\,, \tag{3.53}$$

and since this statistics are constructed with ratios and spacings of order statistics, they are location and scale invariant – ancillary. Once again, the tests are considered at the asymptotic significance level α.

Ratio Test Statistic Proposed in Neves et al. (2006) as a complementary test statistic, it is given by the ratio between the maximum and the mean of the excesses above the random

threshold and its discriminant behavior towards heavy or light tailed distributions is proved to be essentially driven by the sample maximum. It was developed in sight of the differences in the contributions of the maximum to the sum of the k excesses considered in heavy tails, and barely detects small negative values of the EVI. The statistic is given as

$$R_n(k) = \frac{Z_1}{k^{-1}\sum_{i=1}^{k} Z_i} \tag{3.54}$$

and its normalized version has an asymptotic Gumbel behaviour, under the null hypothesis of attraction to the Gumbel max-domain of attraction,

$$R_n^*(k) = R_n(k) - \log(k) \xrightarrow[n\to\infty]{d} T \frown \Lambda \,. \tag{3.55}$$

Given this asymptotic behaviour, the null hypothesis in the two-sided test (3.51) is then rejected if $R_n^*(k) < \mathcal{G}_{\frac{\alpha}{2}}$ or $R_n^*(k) > \mathcal{G}_{1-\frac{\alpha}{2}}$.

The rejection regions for the one-sided tests in (3.52) are given by $R_n^*(k) \leq \mathcal{G}_\alpha$ or $R_n^*(k) \geq \mathcal{G}_{1-\alpha}$ resp. when dealing with attraction to the Weibull or Fréchet alternative max-domain.

Hasofer-Wang Test Statistic Firstly introduced and named after Hasofer and Wang (1992), it was based on the fixed top k order statistics and is a generalization of the Shapiro-Wilk goodness-of-fit statistic. In this article, its power was proved to be superior to that of other tests that had been previously proposed. The asymptotic properties of this statistic were reformulated in Neves and Fraga Alves (2007) with resource to regular variation theory, when considering that k behaves as the intermediate sequence k_n rather than remaining fixed while the sample size increases. It is the most powerful test of the three here presented when studying alternatives in the Weibull domain of attraction. The statistic is given as

$$W_n(k) = k^{-1}\frac{\left(k^{-1}\sum_{i=1}^{k} Z_i\right)^2}{k^{-1}\sum_{i=1}^{k} Z_i^2 - \left(k^{-1}\sum_{i=1}^{k} Z_i\right)^2} \tag{3.56}$$

and it's normalized version has an asymptotic Normal behaviour, under the null hypothesis of attraction to the Gumbel max-domain,

$$W_n^*(k) = \sqrt{\frac{k}{4}}\,(k\,W_n(k) - 1) \xrightarrow[n\to\infty]{d} W \frown \mathcal{N}(0,1) \,. \tag{3.57}$$

The null hypothesis in the two-sided test (3.51) is then rejected if $|W_n^*(k)| \geq z_{1-\frac{\alpha}{2}}$.

For the one-sided tests in (3.52), the rejection regions are given by $W_n^*(k) \geq z_{1-\alpha}$ or $W_n^*(k) \leq z_\alpha$ resp. when dealing with attraction to the Weibull or Fréchet alternative max-domain.

Greenwood Test Statistic Firstly introduced and named after Greenwood (1946), it was re-defined in Neves and Fraga Alves (2007) by multiplying the original statistic by k, and its asymptotic properties reformulated similarly to what was done with the Hasofer-Wang test statistic. Also based on the spacings of high order statistics (on the excesses), it was shown

to be advantageous when testing the presence of heavy-tailed distributions is demanded. Its discriminating behavior towards heavy-tailed distributions accounts for a slightly more powerful test than the Hasofer and Wang's testing procedure in this situations, even though it barely detects small values of the EVI. The statistic is given as

$$Gr_n(k) = \frac{k^{-1}\sum_{i=1}^{k} Z_i^2}{\left(k^{-1}\sum_{i=1}^{k} Z_i\right)^2} \tag{3.58}$$

and it's normalized version has an asymptotic Normal behaviour, under the null hypothesis of attraction to the Gumbel max-domain,

$$Gr_n^*(k) = \sqrt{\frac{k}{4}}\,(Gr_n(k) - 2) \xrightarrow[n\to\infty]{d} G \frown \mathcal{N}(0,1)\ . \tag{3.59}$$

The null hypothesis in the two-sided test (3.51) is then rejected if $|Gr_n^*(k)| \geq z_{1-\frac{\alpha}{2}}$.

For the one-sided tests in (3.52), the rejection regions are given by $Gr_n^*(k) \leq z_\alpha$ or $Gr_n^*(k) \geq z_{1-\alpha}$ resp. when dealing with attraction to the Weibull or Fréchet alternative max-domain.

A different type of test statistic is suggested in Fraga Alves et al. (2016) as a useful tool for either discarding heavy-tailed models or detecting short-tailed models, developed from the *general right endpoint estimator*, the focus of the referred paper. Moreover, it is detached of any external estimation of the EVI, since the general right endpoint estimator upon which it is built does not depend on any estimation of ξ. This testing procedure is shown to detect short-tails better than the Ratio test statistic, and to out-perform the Greenwood statistic for models with $\xi = -1/2$. Let us denote $\hat{x}_g^F$ the general right endpoint estimator, which will be defined in a posterior section, and consider the intermediate sequence k_n as previously described. The statistic is given as

$$G_{n,k} = \frac{\hat{x}_g^F - X_{n-k:n}}{X_{n-k:n} - X_{n-2k:n}} \tag{3.60}$$

and it's normalized version has an asymptotic Gumbel behaviour, under the null hypothesis of attraction to the Gumbel max-domain,

$$G_{n,k}^*(0) = \log(2)\, G_{n,k} - \frac{\log(2k^2)}{2} \xrightarrow[n\to\infty]{d} Z \frown \Lambda\ . \tag{3.61}$$

The rejection region for the two-sided test postulated in (3.51), also at an asymptotic level α, is given by conditions $G_{n,k}^*(0) \leq \mathcal{G}_{\frac{\alpha}{2}}$ or $G_{n,k}^*(0) \geq \mathcal{G}_{1-\frac{\alpha}{2}}$. For testing the one-sided counterparts in (3.52), the rejection regions are defined by $G_{n,k}^*(0) \leq \mathcal{G}_\alpha$ or $G_{n,k}^*(0) \geq \mathcal{G}_{1-\alpha}$ resp. when dealing with attraction to the Weibull or Fréchet alternative max-domain.

Another way of approaching the statistical choice of domain of attraction problem is to analyze if a finite upper bound is acceptable. The test procedure here presented is described in Neves and Pereira (2010) for processes with positive observations and enables us to distinguish light-tailed distribution functions with finite right endpoint from those with infinite endpoint lying in the Gumbel domain. A practical application is shown in Fraga Alves et al. (2016).The

hypothesis for detecting finiteness of the right endpoint for a distribution which may belong to either Weibull or Gumbel domains are

$$H_0 : \; F \in \mathcal{D}_{\mathcal{M}}(G_0),\; x^F = \infty \qquad versus \qquad H_1 : \; F \in \mathcal{D}_{\mathcal{M}}(G_\xi)_{\xi \leq 0},\; x^F < \infty \,. \tag{3.62}$$

The test statistic is given in terms of

$$T_{n,k}^{(1)} = k^{-1} \sum_{i=1}^{k} \frac{X_{(i+1)} - X_{(k+1)} - a_{n,k}}{X_{(1)} - X_{(k+1)}} \tag{3.63}$$

were $a_{n,k}$ is a suitable estimator for the normalizing scaling constant $a\left(\frac{n}{k}\right)$ in (3.50). The authors suggest (pertaining to the moment statistics from Dekkers et al. (1989))

$$a_{n,k} = X_{(k+1)} \, \frac{M_{n,k}^{(1)}}{2} \left(1 - \frac{\left(M_{n,k}^{(1)}\right)^2}{M_{n,k}^{(2)}} \right)^{-1} \tag{3.64}$$

$$M_{n,k}^{(j)} = k^{-1} \sum_{i=1}^{k} \left(\log(X_{(i)}) - \log(X_{(k+1)}) \right)^j , \quad j = 1, 2 \,. \tag{3.65}$$

Under the referred null hypothesis H_0 is valid a normal asymptotic behaviour

$$T_1^* = \sqrt{k} \, \log(k) \, T_{n,k}^{(1)} \xrightarrow[n \to \infty]{d} T \frown \mathcal{N}(0,1) \,. \tag{3.66}$$

Rejection regions at a significance level α are thusly given by $|T_1^*| \geq z_{1-\frac{\alpha}{2}}$. However, this testing procedure is not very sharp for detecting very small negative values of the EVI.

3.2.2 Estimation of the Extreme Value Index

In this section we introduce some existing estimators, both classical and more recent, for the EVI under the semi-parametric setup described. We reinforce the importance of choosing the value of k appropriately as an intermediate sequence, since the inference will be based on the top $k+1$ values above a random threshold $X_{(k+1)}$ from the original n-sized sample. Note that $k+1$ represents merely the portion of the sample's top o.s.'s selected for the estimation, not entailing that all $k+1$ are used in the computation of each estimator. This is, some of the presented estimators select the sample fraction based on the random threshold, but the number of observations used is smaller than $k+1$.

Pickands Estimator Named after its author Pickands (1975), was the first semi-parametric estimator for any real EVI $\xi \in \mathbb{R}$ to be introduced. It is given by the functional form

$$\hat{\xi}_{n,k}^P = \frac{1}{\log(2)} \log \left(\frac{X_{([\frac{k}{4}])} - X_{(2[\frac{k}{4}])}}{X_{(2[\frac{k}{4}])} - X_{(4[\frac{k}{4}])}} \right) , \qquad \frac{k}{4} = 1, 2, \ldots, \left[\frac{n}{4}\right] \tag{3.67}$$

where $[x]$ denotes the integer part of x.

Dekkers and de Haan (1989) studied the asymptotic behaviour of this estimator, proving, among other properties, the strong consistency property we now enunciate. This of course implicates the weak consistency of $\hat{\xi}^P_{n,k}$, which was previously shown by Pickands (1975).

Theorem 3.2.1. *Let $X_1, \ldots, X_n, \ldots$ be an i.i.d. sequence of r.v.'s with d.f. $F \in \mathcal{D}_{\mathcal{M}}(G_\xi)$ for some $\xi \in \mathbb{R}$. If k_n is an intermediate sequence of integers in the sense of* (3.45) *such that $\frac{k_n}{\log(\log(n))} \to \infty$, then*

$$\hat{\xi}^P_{n,k} \xrightarrow[n\to\infty]{a.s.} \xi\,,$$

where $\xrightarrow[n\to\infty]{a.s.}$ stands for almost sure convergence.

The Pickands estimator is shift and scale invariant and its simplicity and applicability to the general case $\xi \in \mathbb{R}$ make it an attractive estimator, though it shows a very high variance and is very dependent on the chosen value of k. This has motivated several authors over the years to procure ways of modifying and generalizing $\hat{\xi}^P_{n,k}$. We refer, as an example of this pursuit, the works of Fraga Alves (1992, 1995), Themido Pereira (1993) and Yun (2002) where a generalization of the Pickands estimator is suggested with the introduction of a tuning or control parameter.

Hill Estimator Introduced by Hill (1975) shortly after the Pickands estimator was introduced, the Hill estimator is conditioned to the case of heavy tails, that is, can only be used for estimating $\xi > 0$. It is given by the functional form

$$\hat{\xi}^H_{n,k} = k^{-1} \sum_{i=1}^{k} \big(\log(X_{(i)}) - \log(X_{(k+1)})\big) =: M^{(1)}_{n,k} \tag{3.68}$$

where $M^{(1)}_{n,k}$ is the moment statistic defined in (3.65).

The study of its asymptotic properties has been performed by various authors and can be found summarized, for example, in de Haan and Ferreira (2006). Similarly to the Pickands estimator, it has been proven that the Hill estimator is strongly consistent, and therefore also weakly consistent.

Theorem 3.2.2. *Let $X_1, \ldots, X_n, \ldots$ be an i.i.d. sequence of r.v.'s with d.f. $F \in \mathcal{D}_{\mathcal{M}}(G_\xi)$ for some $\xi > 0$. If k_n is an intermediate sequence of integers in the sense of* (3.45) *such that $\frac{k_n}{\log(\log(n))} \to \infty$, then*

$$\hat{\xi}^H_{n,k} \xrightarrow[n\to\infty]{a.s.} \xi\,.$$

Unlike the Pickands estimator, $\hat{\xi}^H_{n,k}$ is not shift invariant, although it remains unaffected to changes is scale. Another disadvantage comes from its sensitivity to the rate of growth of the intermediate sequence k_n when n goes to infinity, which can induce a large bias. Still, it is a largely applied estimator for the tail index of d.f.'s on the Fréchet max-domain of attraction (which clearly exposes the necessity of performing pre-tests on the signal of the EVI, has shown in the previous section).

Given its handicaps, there have been several generalizations an adaptations of this estimator proposed by several authors. Some will be presented in detail ahead, but we refer here the

works of Peng (1998), who addressed the problem of finding an unbiased version of the Hill estimator, and Fraga Alves (2001), who proposed a shift-invariant estimator based on $\hat{\xi}^H_{n,k}$.

Negative Hill Estimator Introduced by Falk (1995) in a time when the asymptotic properties of the Maximum Likelihood estimator of the EVI in the parametric POT methodology (in 3.1.2) were only known for the case $\xi > -0.5$, the Negative Hill estimator constitutes an alternative applicable when $\xi < 0$ but intended for the specific cases when $\xi < -0.5$.

Recall that the attraction to a Weibull max-domain means a short, bounded right tail, that is, a finite right endpoint $x^F < \infty$, which can be well estimated for $\xi < -0.5$ simply by the sample maximum. Thus, according to our go-to reference de Haan and Ferreira (2006), under the condition of $F \in \mathcal{D}_{\mathcal{M}}(G_\xi)_{\xi<0}$ we can define the r.v. $\mathbb{X} = \frac{1}{x^F - X}$ (with x^F estimated by $X_{(1)}$) being attracted to the max-domain of $G_{-\xi}$, and as such the Hill estimator in (3.68) can be used on the transformed variable $\mathbb{X}$. This gives us the functional form of the Negative Hill estimator

$$\hat{\xi}^{NH}_{n,k} = k^{-1} \sum_{i=1}^{k-1} \log(X_{(1)} - X_{(i+1)}) - \log(X_{(1)} - X_{(k+1)}) \,. \tag{3.69}$$

This estimator is shift as well as scale invariant, unlike the merely scale invariant Hill estimator. For the intended application values of the EVI $\xi < -0.5$ the Negative Hill estimator is proven to be consistent (see once again proof in de Haan and Ferreira (2006)).

Theorem 3.2.3. *Let $X_1, \ldots, X_n, \ldots$ be an i.i.d. sequence of r.v.'s with d.f. $F \in \mathcal{D}_{\mathcal{M}}(G_\xi)$ for some $\xi < -0.5$. If k_n is an intermediate sequence of integers in the sense of* (3.45) *such that $\frac{k^\eta}{\log(n)} \to \infty$ for $\eta \to 0$, then*

$$\hat{\xi}^{NH}_{n,k} \xrightarrow[n\to\infty]{p} \xi$$

where $\xrightarrow{p}$ stands for convergence in probability.

Generalized Hill Estimator An adaptation of the Hill estimator for the general case $\xi \in \mathbb{R}$ was introduced by Beirlant et al. (1996), derived directly from $\hat{\xi}^H_{n,k}$ in (3.68) as shows its functional expression

$$\hat{\xi}^{GH}_{n,k} = \hat{\xi}^H_{n,k} + k^{-1} \sum_{i=1}^{k} \left(\log(\hat{\xi}^H_{n,i}) - \log(\hat{\xi}^H_{n,k}) \right) . \tag{3.70}$$

This estimator is invariant to modifications of scale but, as the Hill estimator, not to modifications on the location. It was also proven to be consistent in the full range $\xi \in \mathbb{R}$. A study of the asymptotic properties of the Generalized Hill estimator can be found in Beirlant et al. (2005).

Moment Estimator Developed as a generalization of the Hill estimator applicable to any domain of attraction (i.e. for $\xi \in \mathbb{R}$), the Moment estimator was presented in the work of Dekkers et al. (1989).

Recall the definition of the moment statistics $M_{n,k}^{(j)}$ in (3.65), developed in that same paper. Consider as well

$$\hat{\xi}_{n,k}^{+} = M_{n,k}^{(1)} = \hat{\xi}_{n,k}^{H} \qquad \text{and} \qquad \hat{\xi}_{n,k}^{-} = 1 - \frac{1}{2}\left(1 - \frac{\left(M_{n,k}^{(1)}\right)^2}{M_{n,k}^{(2)}}\right)^{-1} . \tag{3.71}$$

The Moment estimator for $\xi = \xi^+ + \xi^-$ (where the notations signify $\xi^+ = \max\{0, \xi\}$ and $\xi^- = \min\{0, \xi\}$) is then composed of 2 parts: $\hat{\xi}_{n,k}^{+}$, the Hill estimator, valid for when $\xi > 0$, and $\hat{\xi}_{n,k}^{-}$, which is called the *Negative Moment* estimator, valid for when $\xi < 0$. Hence, the Moment estimator is given as

$$\hat{\xi}_{n,k}^{M} = \hat{\xi}_{n,k}^{+} + \hat{\xi}_{n,k}^{-} . \tag{3.72}$$

The property of scale but not shift invariance of the Hill estimator is preserved by this Moment estimator, as it is the property of strong (and consequently weak) consistency.

Theorem 3.2.4. *Let $X_1, \ldots, X_n, \ldots$ be an i.i.d. sequence of r.v.'s with d.f. $F \in \mathcal{D}_{\mathcal{M}}(G_\xi)$ for some $\xi \in \mathbb{R}$. If k_n is an intermediate sequence of integers in the sense of* (3.45) *such that $\frac{k_n}{(\log(n))^\eta} \to \infty$ for some $\eta > 0$, then*

$$\hat{\xi}_{n,k}^{M} \xrightarrow[n\to\infty]{a.s.} \xi .$$

Negative Moment Estimator As shown above, the Negative Moment estimator for $\xi < 0$ is simply given by

$$\hat{\xi}_{n,k}^{NM} = \hat{\xi}_{n,k}^{-} = 1 - \frac{1}{2}\left(1 - \frac{\left(M_{n,k}^{(1)}\right)^2}{M_{n,k}^{(2)}}\right)^{-1} . \tag{3.73}$$

This estimator is consistent for the estimation of $\xi^- = \min\{0, \xi\}$ in the whole $\mathcal{D}_{\mathcal{M}}(G_\xi)_{\xi<0}$ (Gomes et al. (2013b)).

Based on this estimator, Caeiro and Gomes (2010) suggested a class of consistent estimators for $\xi < 0$, generalizing at once both the Moment and Negative Moment estimators. With the introduction of a *tuning parameter* θ, the family of estimators is given by

$$\hat{\xi}_{n,k}^{NM(\theta)} = \theta\, \hat{\xi}_{n,k}^{H} + \hat{\xi}_{n,k}^{NM} , \quad \theta \in \mathbb{R}. \tag{3.74}$$

Note that for $\theta = 0$ we have the Negative Moment estimator, and for $\theta = 1$ the Moment estimator. With the appropriate choice of k and of the tuning parameter, this estimator can have a smaller asymptotic bias than $\hat{\xi}_{n,k}^{M}$.

Mixed Moment Estimator This estimator for the general scenario $\xi \in \mathbb{R}$ was developed by Fraga Alves et al. (2009b), based in a combination of Theorem 2.6.1 and Theorem 2.6.2 of de Haan (1970). The Mixed Moment estimator is then given by

$$\hat{\xi}_{n,k}^{MM} = \frac{\hat{\varphi}_n(k) - 1}{1 + 2\,\min\{\hat{\varphi}_n(k) - 1, 0\}} \tag{3.75}$$

where

$$\hat{\varphi}_n(k) = \frac{M_{n,k}^{(1)} - L_{n,k}^{(1)}}{\left(L_{n,k}^{(1)}\right)^2} \qquad \text{with} \qquad L_{n,k}^{(1)} = k^{-1} \sum_{i=1}^{k} \left(1 - \frac{X_{(k+1)}}{X_{(i)}}\right)$$

and $M_{n,k}^{(1)}$ the first moment statistic defined in (3.65).

The Mixed Moment estimator is consistent for any real value of the EVI (Corollary 2.1 of Fraga Alves et al. (2009b)), and the simple explicit functional form is one of its appealing qualities. However, it is not location invariant and hence some alternatives that stay unchanged in the face of sifts in location are presented in that same paper, as we will see ahead.

Location Invariant Moment Estimator As stated above, the Moment estimator presented by Dekkers et al. (1989), and consequently the Negative Moment estimator, are only scale invariant, being vulnerable to shifts in location. So, Ferreira et al. (2003) introduced a modification to $\hat{\xi}_{n,k}^{NM}$ that made it also shift invariant. However, this new estimator can only be used in the whole $\mathcal{D}_{\mathcal{M}}(G_\xi)_{\xi<0}$, contrary to the Moment estimator that is applicable to any real EVI.

The main functional difference in this Location Invariant Moment estimator is considering, instead of the moment statistics defined in (3.65),

$$N_{n,k}^{(j)} = k^{-1} \sum_{i=1}^{k} \left(X_{(i)} - X_{(k+1)}\right)^j, \quad j = 1, 2\,. \tag{3.76}$$

As such, this estimator's functional form is, analogously to the Negative Moment estimator,

$$\hat{\xi}_{n,k}^{IM} = 1 - \frac{1}{2}\left(1 - \frac{\left(N_{n,k}^{(1)}\right)^2}{N_{n,k}^{(2)}}\right)^{-1}. \tag{3.77}$$

However, we were not able to find in the literature any reference for the explicit asymptotic properties of this estimator, and as such it will be left out of the study performed in the upcoming section 3.2.4.

Peaks Over Random Threshold Estimators Many of the estimators mentioned up to this point, based on the excesses of the log-observations, are not invariant to changes in location, such as the Hill, Moment or Mixed Moment estimators.

It is statistically interesting to obtain sift invariant estimators, since applying non-invariant estimators in situations where the location parameter suffers a change can lead to serious errors in any estimation performed.

The most widely applied approach to this problem is the one suggested by Araújo Santos et al. (2006) and commonly denominated Peaks Over Random Threshold (PORT) methodology. This is based on the introduction of a *tuning parameter* $q \in (0, 1)$ that controls the threshold above which the excesses will be used. As such, the functional forms of the

estimators remain the same, but are used on the observations of the sample of excesses

$$\left(X_{n:n} - X_{n_q:n}, X_{n-1:n} - X_{n_q:n}, \ldots, X_{n_q+1:n} - X_{n_q:n}\right)$$

where $n_q := [n \cdot q] + 1$.

The resulting sift invariant estimators maintain similar properties to those of the original estimator, such as the property of consistency, given the very important right choice of the tuning parameter q and of the level k.

An application of this methodology has been performed on the Hill and Moment estimators in the referred paper of Araújo Santos et al. (2006), and more recently extended to the Mixed Moment estimator in Fraga Alves et al. (2009b). Respective studies of the asymptotic properties can be found in the mentioned publications.

The notation used for these estimators is $\hat{\xi}_{n,k}^{E(q)}$, where E stands for the original estimator chosen. In the Case Study of this dissertation, exposed in the next Chapter, will be considered only the PORT-Moment and PORT-Mixed Moment estimators, and as such in the notation above we have $E = M$, MM.

POT-ML Estimator The estimator we now present is different and more complex than the ones listed above, and was introduced by Smith (1987).

Being that the class of functions in the GEV domain of attraction, for any EVI, can't be defined by a finite number of parameters, as explained in de Haan and Ferreira (2006), it is not possible to find a Maximum Likelihood estimator for ξ in this class.

However, considering the excesses above a random threshold Z_i, as defined in (3.53) in the previous section, these are approximately the k top o.s.'s associated with a sample of size k from a Generalized Pareto distribution. Solving the ML equations associated with this approximation, as can be seen in Davison (1984) yields the explicit form of an estimator for the EVI usually noted as ML

$$\hat{\xi}_{n,k}^{ML} = k^{-1} \sum_{i=1}^{k} \log(1 + \hat{\alpha}\, Z_i)\,. \tag{3.78}$$

where $\hat{\alpha}$ is the implicit ML estimator of the unknown scale parameter.

This estimator too is consistent under the First Order Condition and the appropriate choice of an intermediate sequence k_n in the sense of (3.45). More on its asymptotic properties can be found in Drees et al. (2004) and Zhou (2009).

3.2.3 Estimation of Other Relevant Extreme Value Indicators

As was seen before regarding the parametric estimation, there are other interesting parameters of extreme values we wish to infer about. We will now show how to estimate said parameters in the semi-parametric setup, assuming the EVI has been estimated as $\hat{\xi}$ by at least one of the estimators suggested in the previous section.

We start from the basic assumption in semi-parametric inference: $F \in \mathcal{D}_{\mathcal{M}}(G_\xi)$ for some $\xi \in \mathbb{R}$. We know from Theorem 2.3.1 that this is equivalent to the First Order Condition in (2.20), which allows us to approximate

$$U(tx) \approx U(t) + a(t)\frac{x^\xi - 1}{\xi}$$

under the conditions of the referred Theorem. A simple variable change gives us the equivalent expression

$$U(x) \approx U(t) + a(t)\frac{\left(\frac{x}{t}\right)^\xi - 1}{\xi}\,. \tag{3.79}$$

Recall the form of $U(.)$, the tail quantile function in Definition 2.3.1, that allows us to express the **Extremal Quantiles** as $\chi_p := U\left(\frac{1}{p}\right)$ when p is a close enough to 0 value: theoretically, $p \equiv p_n \to 0$ with $p_n \ll \frac{k}{n}$ as $n \to \infty$. Taking p in this conditions and $t = \frac{n}{k} \to \infty$ with k and intermediate sequence in the sense of (3.45), we can approximate the extremal quantile through (3.79) as

$$U\left(\frac{1}{p}\right) \approx U\left(\frac{n}{k}\right) + a\left(\frac{n}{k}\right)\frac{\left(\frac{k}{np}\right)^\xi - 1}{\xi}\,. \tag{3.80}$$

Theorem 2.3.1 also suggests that a possible choice for the **Location Attraction Coefficient** is $b(t) = U(t)$. Since we are taking $t = \frac{n}{k}$ and given that the we estimate the d.f. F by its empirical counterpart F_n (3.48), we can estimate

$$\hat{b}\left(\frac{n}{k}\right) = \hat{U}\left(\frac{n}{k}\right) = \hat{F}^{\leftarrow}\left(1 - \frac{k}{n}\right) = F_n^{\leftarrow}\left(1 - \frac{k}{n}\right) = X_{n-k:n} = X_{(k+1)}\,. \tag{3.81}$$

This means the random threshold $X_{n-k:n}$ is the estimator of the location attraction coefficient.

As for the **Scale Attraction Coefficient**, we have shown in (3.49) it can be given by $a\left(\frac{n}{k}\right)$. A consistent estimator for this quantity, i.e., an estimator that satisfies $\frac{\hat{a}\left(\frac{n}{k}\right)}{a\left(\frac{n}{k}\right)} \xrightarrow[n\to\infty]{p} 1$ as proven by Theorem 4.2.1 of de Haan and Ferreira (2006), is

$$\hat{a}\left(\frac{n}{k}\right) = X_{(k+1)}\, M_{n,k}^{(1)}\left(1 - \hat{\xi}_{n,k}^{-}\right)\,, \tag{3.82}$$

where $M_{n,k}^{(1)}$ is the first moment statistic define in (3.65) and $\hat{\xi}_{n,k}^{-}$ is defined in (3.71) as the negative part of the Moment estimator. When the case is that of $\xi > 0$, we have $\hat{\xi}_{n,k}^{-} \xrightarrow[n\to\infty]{p} 0$, which yields the simpler form of the scale coefficient estimator

$$\hat{a}\left(\frac{n}{k}\right) = X_{(k+1)}\, M_{n,k}^{(1)}\,. \tag{3.83}$$

Having these estimators for the attraction coefficients, and assuming $\hat{\xi}$ is one of the EVI's consistent estimators considered before, we can now explicitly write the **Extremal Quantile**

estimator, when $\xi \neq 0$, from expression (3.80) as

$$\widehat{\chi_p} = \hat{U}\left(\frac{1}{p}\right) = X_{(k+1)} + \hat{a}\left(\frac{n}{k}\right)\frac{\left(\frac{k}{n\,p}\right)^{\hat{\xi}} - 1}{\hat{\xi}} \tag{3.84}$$

which reduces to

$$\widehat{\chi_p} = \hat{U}\left(\frac{1}{p}\right) = X_{(k+1)} + \hat{a}\left(\frac{n}{k}\right)\log\left(\frac{k}{n\,p}\right) \tag{3.85}$$

when $\xi = 0$, with the scale coefficient estimator given in (3.82). However, Weissman (1978) suggested a simpler form of this estimator when dealing with heavy-tailed distributions, i.e., for the case $\xi > 0$, which became known as the Weissman estimator

$$\widehat{\chi_p} = \hat{U}\left(\frac{1}{p}\right) = X_{(k+1)}\left(\frac{k}{n\,p}\right)^{\hat{\xi}},$$

with $\hat{\xi}$ here the Hill estimator. Recall that **Return Levels** as presented in the parametric approach are simply particular cases of these extremal quantiles and can be estimated through these same estimators.

For estimating the **Exceedance Probability** of a level x, denoted as $p = 1 - F(x)$, we will make use of the approximation of the tail function $\overline{F}$ expressed in (3.50). The high levels x to be considered must theoretically satisfy the condition $x \equiv x_n$ such that $\overline{F}(x_n) =: p_n \to 0$ as $n \to \infty$. Then, given the approximation (3.50) and the expression of GP's d.f. H_ξ, it can be proven that a consistent estimator for the exceedance probability p_n, i.e. an estimator that satisfies $\frac{\hat{p}_n}{p_n} \xrightarrow[n\to\infty]{p} 1$, when $\xi \neq 0$, is

$$\hat{p}_n = \hat{\overline{F}}(x_n) = \frac{k}{n}\left\{\max\left(0,\, 1 + \hat{\xi}\frac{x_n - X_{(k+1)}}{\hat{a}\left(\frac{n}{k}\right)}\right)\right\}^{-1/\hat{\xi}}, \tag{3.86}$$

and for $\xi = 0$, by continuity arguments,

$$\hat{p}_n = \hat{\overline{F}}(x_n) = \frac{k}{n}\exp\left\{-\frac{x_n - X_{(k+1)}}{\hat{a}\left(\frac{n}{k}\right)}\right\}. \tag{3.87}$$

Once again, it is possible to obtain a simpler estimator when dealing with heavy-tailed distributions, i.e., for the case $\xi > 0$,

$$\hat{p}_n = \hat{\overline{F}}(x_n) = \frac{k}{n}\left\{\frac{x_n}{X_{(k+1)}}\right\}^{-1/\hat{\xi}}, \tag{3.88}$$

motivated by the expression of the simpler scale attraction coefficient estimator in (3.83) – more details on the simplification can be found in Gomes et al. (2013a).

Recall that **Return Periods**, as presented in the parametric approach, can be calculated as the inverse of the exceedance probability of a given return level u, and as such estimated with resource to the exceedance probability estimators here constructed.

The **Right Endpoint**, denoted x^F, is also a very important indicator we wish to estimate, particularly when in a Weibull-max domain where this quantity is finite. In Definition 2.3.1 we

identified $x^F = U(\infty)$. We will now combine this relation with the extremal quantile estimator presented in (3.84) to obtain an estimator of the finite right endpoint when $\xi < 0$. Taking $p = 0$, the expression of the estimator is

$$\widehat{x^F} = \hat{U}(\infty) = X_{(k+1)} - \frac{\hat{a}\left(\frac{n}{k}\right)}{\hat{\xi}} \tag{3.89}$$

where $\hat{a}\left(\frac{n}{k}\right)$ is defined in (3.82) and $\hat{\xi}$ is any of the estimators for the EVI considered in the previous section. In practice, there isn't always guarantee that the obtained estimate will be admissible, since an estimate for the right endpoint smaller than the sample maximum is clearly wrong. Therefore, it is usual to consider the more general form of the right endpoint estimator as

$$\widehat{x^F} = \max\left\{X_{(1)},\, X_{(k+1)} - \frac{\hat{a}\left(\frac{n}{k}\right)}{\hat{\xi}}\right\}. \tag{3.90}$$

As we can observe, all this estimators depend on the previous external estimation of the tail parameter ξ. In Fraga Alves and Neves (2014), the authors present a new estimator for finite right endpoint in the Gumbel domain, being later extended for any light-tailed distribution function belonging to some max-domain of attraction in Fraga Alves et al. (2016). It was named *General Right Endpoint Estimator*, and it does not require any previous estimation of the EVI (which is supposed to be non-positive). This estimator has been referred in this dissertation at the point of the introduction in Section 3.2.1 of a testing procedure for the EVI sign based on the $G_{n,k}$ statistic in (3.60), which is constructed with resource to this general right endpoint estimator. This estimator has the advantage of always yielding admissible estimates for the right endpoint, that is, its values are always greater than the sample maximum $X_{(1)}$. It is given by the functional form

$$\widehat{x^F}_g := X_{(1)} + X_{(k+1)} - \frac{1}{\log(2)} \sum_{i=0}^{k-1} \log\left(1 + \frac{1}{k+i}\right) X_{(k+i+1)}\,. \tag{3.91}$$

An alternative and equivalent form can be obtained by defining the new quantities $a_{i,k} = \frac{\log\left(\frac{k+i+1}{k+ik}\right)}{\log(2)}$, yeilding the expression

$$\widehat{x^F}_g := X_{(1)} + \sum_{i=0}^{k-1} a_{i,k}\left(X_{(k+1)} - X_{(k+i+1)}\right) \qquad \text{with} \qquad \sum_{i=0}^{k-1} a_{i,k} = 1\,.$$

This estimator is consistent, as proven in Fraga Alves et al. (2016) and it has a broader spectrum of application than the usual alternatives.

3.2.4 Estimators' Asymptotic Properties

We have seen that all the EVI's estimators presented in section 3.2.2 above are consistent, a property depending only on the behaviour of the intermediate sequence k_n when the sample size n grows towards infinity (with the referred exception of the Location Invariant Moment estimator, for which we could not find proof for asymptotic results).

Moreover, we will now show that it is possible to guarantee the asymptotical normality of said

estimators, conditional on the validity of the Second Order Condition, ascertained in Definition 2.3.3, as thusly construct approximated confidence intervals for the tail index ξ.

Let us denote $\hat{\xi}^E_{n,k}$ an arbitrary aforementioned EVI estimator from section 3.2.2, being referred respectively by considering $E \in \mathbb{E} = \{P, H, NH, GH, M, NM, MM, M(q), MM(q), ML\}$, defined for the corresponding applicable values of ξ.

Theorem 3.2.5. *Assuming that the Second Order Condition is satisfied by the d.f. F and that $k \equiv k_n$ is an intermediate sequence in the sense of (3.45), if*

$$\lim_{n\to\infty} \sqrt{k}\, A\left(\frac{n}{k}\right) = \lambda$$

for a finite value of λ, then there exists $B_E \in \mathbb{R}$ and $\sigma_E > 0$ such that

$$\sqrt{k}\left(\hat{\xi}^E_{n,k} - \xi\right) \xrightarrow[n\to\infty]{d} Z \backsim \mathcal{N}\left(\lambda B_E, \sigma_E^2\right) . \tag{3.92}$$

Some particular extra mild conditions may be needed for some specific estimators $\hat{\xi}^E_{n,k}$.

This theorem, from de Haan and Ferreira (2006), is the base for the construction of the desired CI's for ξ. The quantities $B_E \in \mathbb{R}$ and $\sigma_E > 0$ required for the result represent resp. the *asymptotical bias* and *asymptotical variance* of $\hat{\xi}^E_{n,k}$, and general expressions can be found in de Haan and Ferreira (2006) and Beirlant et al. (2004). These need to be estimated, and such task isn't always easy. When the point is simply the construction of this CI's, there is the recommendation from de Haan and Ferreira (2006) to assume $\lambda = 0$, so that the limiting normal distribution is centered around 0. This assumption, made for simplicity reasons, has the upside of helping us "dodge" the estimation of the bias B_E, which usually depends on second-order parameters like ρ that are very complex to estimate.

Thereby, the aproximated $(1-\alpha)\times 100\%$ confidence intervals for the EVI based on $\hat{\xi}^E_{n,k}$ are given by

$$\hat{\xi}^E_{n,k} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_E^2}{k}} \tag{3.93}$$

where $z_{1-\frac{\alpha}{2}}$ denotes the standard Normal distributions's $(1-\frac{\alpha}{2})$-quantile. Given that the asymptotic variance of each estimator is dependent on the unknown value of ξ, we will estimate this quantity according to the following expressions, by replacing the EVI by its respective estimate and consequently using $\hat{\sigma}_E$ to replace σ_E in the construction of the CI's.

We stress here than some of these expressions for the variances are only valid if the particular extra conditions for each specific estimator are satisfied. For confirmation of these asymptotic variances we refer the works: de Haan and Ferreira (2006) – for the Pickands, Hill, Moment and Negative Hill estimators; Beirlant et al. (2005) – for the Generalized Hill estimator; Fraga Alves et al. (2009a) – for the Mixed Moment and PORT-Mixed Moment estimators; Araújo Santos et al. (2006) – for the PORT-Moment estimator; Gomes et al. (2013b) – for the Negative Moment estimator; Gomes et al. (2013a) – for the POT-ML estimator.

$$
\begin{aligned}
&\sigma_P^2 = \begin{cases} \dfrac{\xi^2\left(2^{2\xi+1}+1\right)}{(\log(2))^2\left(2^\xi-1\right)^2}, & \xi \neq 0 \\ \dfrac{3}{(\log(2))^4}, & \xi = 0 \end{cases}; \\
&\sigma_H^2 = \xi^2, \quad \xi > 0\,; \\
&\sigma_{NH}^2 = \xi^2, \quad -1 < \xi < -0.5\,; \\
&\sigma_{GH}^2 = \begin{cases} \xi^2+1, & \xi \geq 0 \\ \dfrac{(1-\xi)\left(1+\xi+2\xi^2\right)}{(1-2\xi)}, & \xi < 0 \end{cases}; \\
&\sigma_M^2 = \begin{cases} \xi^2+1, & \xi \geq 0 \\ \dfrac{(1-\xi)^2(1-2\xi)\left(1-\xi+6\xi^2\right)}{(1-3\xi)(1-4\xi)}, & \xi < 0 \end{cases}; \\
&\sigma_{NM}^2 = \sigma_M^2, \quad \xi < 0\,; \\
&\sigma_{MM}^2 = \begin{cases} (\xi+1)^2, & \xi \geq 0 \\ (1-2\xi)^4\,\dfrac{(1-\xi)^2\left(1-\xi+6\xi^2\right)}{(1-3\xi)(1-4\xi)(1-2\xi)^3}, & \xi < 0 \end{cases}; \\
&\sigma_{M(q)}^2 = \sigma_M^2\,; \\
&\sigma_{MM(q)}^2 = \sigma_{MM}^2\,; \\
&\sigma_{ML}^2 = (\xi+1)^2\,.
\end{aligned}
\tag{3.94}
$$

Theorem 3.2.5 further allows us to construct approximated CI's for the right endpoint of a short-tailed distribution F, always bearing in mind that this endpoint can never be estimated lower than the sample maximum – this leads us to define $X_{(1)}$ as the lower limit of the CI and as such we cannot know its exact confidence level, only that it will be less than $(1-\alpha)\times 100\%$. Having $\widehat{x^F}_E$ the right endpoint estimator in (3.89) associated with $\hat{\xi}^E_{n,k}$ and $\hat{a}\left(\frac{n}{k}\right)$ the scale attraction coefficient estimator in (3.82), the interval is then given as

$$
X_{(1)} < x^F < \widehat{x^F}_E + z_{1-\alpha}\frac{\hat{a}\left(\frac{n}{k}\right)}{\left(\hat{\xi}^E_{n,k}\right)^2}\sqrt{\frac{\sigma_E^2}{k}}\,. \tag{3.95}
$$

3.2.5 Determining the Tail Sample Fraction

Through this section on semi-parametric inference it has been recurrently emphasised how important it is to appropriately choose $k \equiv k_n$, the tail sample fraction, so that the suggested estimators have statistically appealing properties, such as consistency. To that purpose, we have been assuming without exception the most basic theoretical condition: k_n must be an intermediate sequence, in the sense of (3.45). But, in practice how to discern which k induced thresholds $X_{(k+1)}$ are convenient for the semi-parametric procedures?

When the sample size n is finite (which is always the case in practical applications), the semi-

parametric EVI estimators presented in section 3.2.2 exhibit different characteristics depending on the choice of the tail sample fraction: small values of k mean that not many order observations will be used in the estimation (because, as said before, a tail sample fraction of k means, for the aforementioned estimators, that the number of o.s.'s used for computing the estimates is not larger than $k+1$), permitting a small bias of the estimators but also inducing a large variance; too high values of k mean a larger number of o.s.'s will be used in the estimators computation, some of which may not even be converging to the expected underlying limiting distribution, which leads to more of the variability in the data covered, in other words, smaller variance of the estimators, but also a more significantly large bias. As such, the choice for the value of k must aim at a bias-variance tradeoff.

Since the emergence of the semi-parametric methodology in extremes, several methods for choosing the tail sample fraction selected for the estimation have been proposed in the literature, such as the specific criterion suggested by Pickands (1975) in his pioneering article. In this dissertation we will adopt an heuristic procedure recently suggested by Henriques-Rodrigues et al. (2011), without overlooking that, as any heuristic methodology, it must be applied with caution.

Henriques-Rodrigues et al. (2011) suggested a "distance" measurement that allows us to choose the value of k for which most of the estimators are agreeing. The procedure then gives us the hopefully optimal sample fraction for the EVI's estimation as

$$k^{opt} = \arg\min_{k} \sum_{(E,J)\in\mathbb{E}:E\neq J} \left(\hat{\xi}^{E}_{n,k} - \hat{\xi}^{J}_{n,k}\right)^2 , \tag{3.96}$$

where $\hat{\xi}^{E}_{n,k}$ represents one of the estimators of ξ from the set $\mathbb{E}$ presented section 3.2.2. This heuristic translates an expectation for the existence of at least one region (in k) where all of the estimators (or at least most of them) will be consistent with each other, that is, close in value. The authors defend that the convergence in distribution expressed in Theorem 3.2.5 remains valid if we replace k with the k^{opt} resulting from this process.

We can also choose to adapt and apply the heuristic to be used for exceedance probability or right endpoint estimators, since the optimal tail sample fraction for the EVI estimation does not always guarantee the best results for the estimation of other indicators. But even further caution is advisable in this cases, especially for the right endpoint estimators defined in (3.90), because even for a bounded-tail underlying distribution, the estimates are not always "admissable". Imagine, for instance, that there is a region k^* where all or almost all of the estimators for the right endpoint lie below the sample maximum, meaning $\widehat{x^F}_E = X_{(1)}$ for E said estimators. Mathematically, they all have the same value, so the squared distance between each pair will be 0 and as such the suggested heuristic will certainly select k^* as the (or one of the) desired k^{opt}. Thus, this procedure should not be applied blindly and it may be necessary to restrict the values of k searched to avoid this problematic areas. A helpful technique is to plot the series of squared distances against the values of k for a visualization of the best regions, and then compare it against the plot of $\hat{\xi}^{E}_{n,k}$ in that region to ascertain if the selected k^{opt} is an admissible choice.

Chapter 4

Case Study – Record Times of Apnea of Female Competitive Freedivers

In order to apply the Extreme Value Theory methodologies so far exposed in this dissertation it was considered a database of female freediving records, where the focus lies on the largest (most extreme and rare) observations.

Freediving is an activity that dates back to the beginning of humanity itself, and holding our breaths when submerged is an instinct that every infant is born with. However, diving in apnea, as introduced in Chapter 1, has only been an international competitive sport since the 1960's, and the safer regulated competitions organized by AIDA had its starting point as late as 1992. Static Apnea is, as indicated before, the only modality of the freediving competitive world that regards apnea time instead of diving distance. Recall that SA consists in the maximum breath-holding time while floating on the surface of the water or standing on the bottom of a pool with every airway immersed. As such, for each athlete it is recorded how long they stay immersed under this conditions. Afterwards, penalties may be applied to the scoring of competitors, but the time registered (in minutes and seconds) is not altered.

In this dissertation we will concern ourselves with the maximum SA times of female competitive freedivers, having in sight the inference of parameters such as the maximum apnea time statistically possible given the current state-of-the-art or the probability that the current record – Natalia Molchanova's 9 minutes and 2 seconds breath-hold – will be beaten. Physiological arguments will not have any more weight than the common sense in our conclusions, and we do not aim to study the effect of the most important external factors influencing the athletes' achievements.

The original data was retrieved on November 3rd 2015 from AIDA's online available rankings. Since it was still possible that further records were registered relating to the year of 2015, it was chosen to omit all the records set in 2015 from the analysis. Furthermore, AIDA's rankings have only been registered as back as 1999, and we have no information regarding the way the records were registered, which may lead to inference errors beyond our control. Also, as was mentioned before, the marks were measured in minutes and seconds and thus the first step in the data treatment was converting every record to a time scale of seconds.

All the methodology presented in the Extreme Values context had the underlying basic assumption of independence of the observations. However, the rankings that comprise our database include several records for the same diver, for several competitions in different years the athlete had participated on. To demonstrate this point, and the implicit lack of independence in this data set, consider the following plot which shows the best record – the maximum apnea time – recorded each year from 1999 to 2014, labeled according to the athlete who achieved the mark.

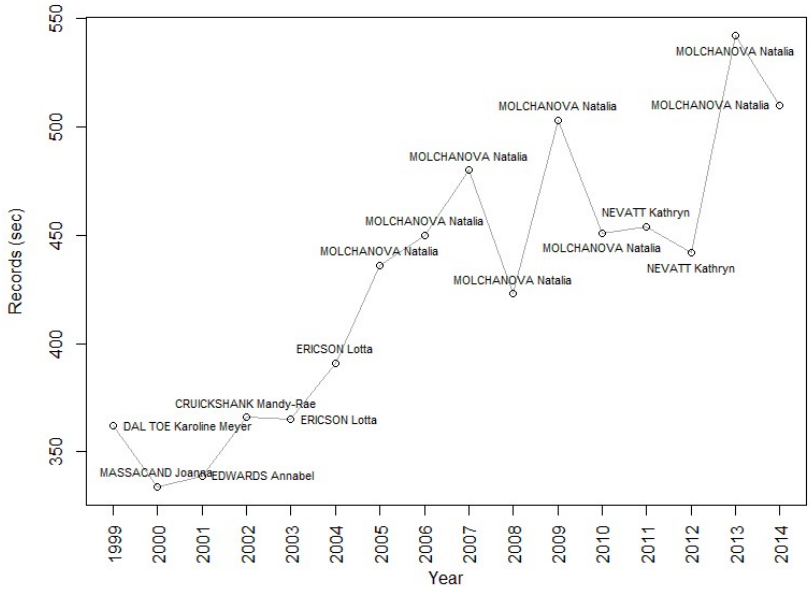

Figure 4.1: Female's SA best annual records

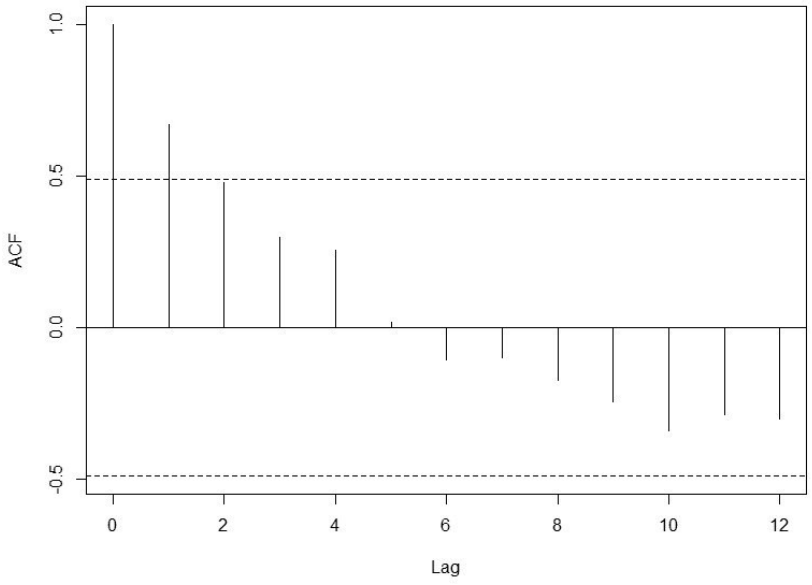

Figure 4.2: ACF of the female's SA best annual records

In the first plot it is striking that the same freediver holds several annual best records, especially the late world recordist Natalia Molchanova, who holds 8 best annual records since 2005. It is then clear that these observations cannot be in any way independent. To support this claim, it was also shown above the plot of the empirical autocorrelation function (ACF) of these 16 observations, where it is clear that correlation exists between the annual bests of two consecutive years.

Given this information, it is safe to assume there is dependence in the complete database at our disposal as well. Therefore, it was selected the single very best performance available for each female freediver, that is, the maximum apnea time registered for each athlete between 1999 and 2014 regardless of the year it happened. This means we will be working only with the most extreme observation for each diver. After this selection, our data is distributed as follows:

Table 4.1: Number of female SA freedivers' individual best records by year.

Year	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**
Count	8	8	12	50	49	63	58	70
Year	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
Count	54	74	71	63	74	125	129	134

As we can see from Table 4.1, a very small number of freedivers had their personal record set in one of the first three years 1999, 2000 and 2001, specially when compared with the number of records set in the most recent years. We suspect there was a change in policy of registering the records motivating this accentuated difference, perhaps allied to a smaller adhesion to the sport in the earlier years. Be that as it may, another censure was applied to the data set: we will only work with female's static apnea best personal records set between 2002 and 2014.

Another question arisen from the fact that every performance from every freediver in an event was registered in the original rankings, which means the existence of many very small values, and as such not all records can be considered as competitive (any human being is able to hold their breath for even a second or two). Even in the censured database where only the best of each diver appears, still many low values were observed. So some criteria had to be employed to select the records worthy of being considered competitive.

To decide on the criteria for this new censuring on the data, two complementary approaches were followed: we turned to the empirical knowledge of the freediving community online, trying to better understand above which level was an apnea time record considered competitive, and on the other hand we plotted all the personal records set from 2002 to 2014 against the year when they took place, to visually separate the smallest records from the median and larger ones.

According to Engineering Sport (2016), it is considered plausible under the right circumstances "the average person being able to hold their breath underwater for around 2 minutes". As we can see in Figure 4.3, there is a reasonable amount of records that fall below the 2 minute mark, so these were all censured out of the data set.

Freedive UK (2016) states that "there is no easy route to a 4+ minute breath-hold". We can see from Figure 4.3 that many of our data fall beyond this mark, but since we presume most of our records belong to professional freedivers we find this to be expectable. We can think of observations above 4 minutes to represent the best of the best records, as we will see ahead.

Finally, in Immersion Freediving (2016) the opinion is that the average student of the modality "does a 2.5-3.5 minute breath hold". Since we already discarded marks below 2 minutes, and given the dispersion in the plot of the records, we settle for ultimately considering that best records above 3 minutes – 180 seconds – are competitive marks.

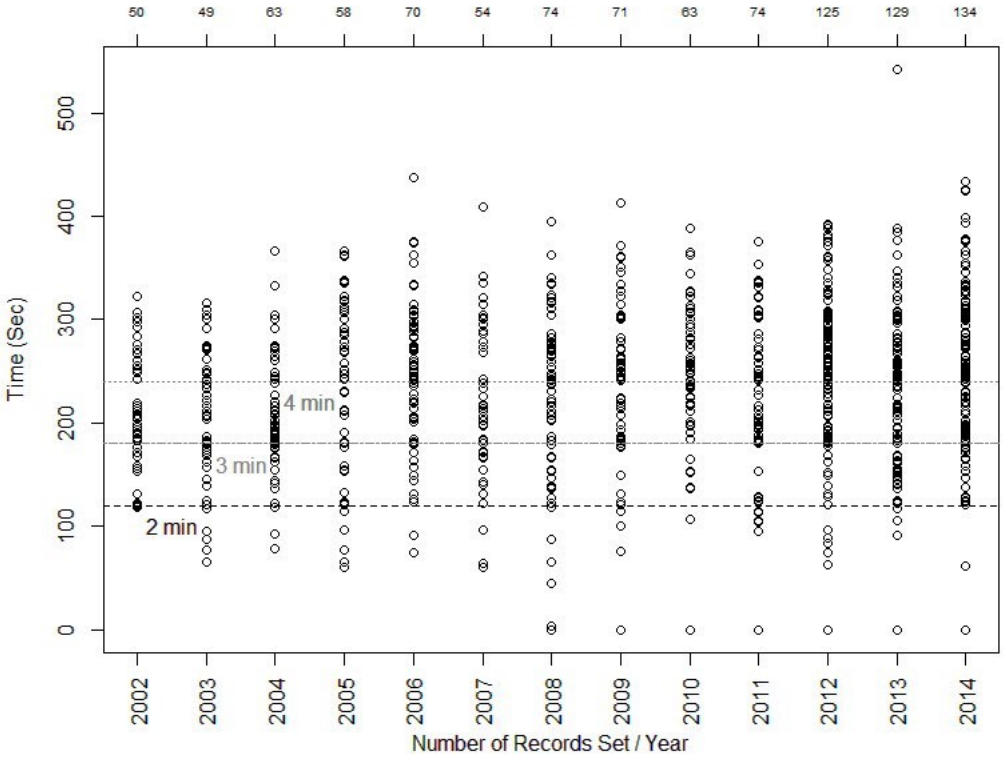

Figure 4.3: Female SA freedivers' individual best records by year

At last, the final data set we will be analysing from now on consists on the Static Apnea female freedivers' personal best records over 180 seconds set between 2002 and 2014, totalling on 795 observations.

Note that we can also observe in Figure 4.1 some positive evolution in the sense that the best yearly marks seem to follow an upward trend. So it will be wise to make a stationarity analysis to determine if this trend prevails on the data set at hand.

Let us now present the complete analysis performed with this data set, following the statistical procedures presented in Chapter 3, within the theoretical mainframe set in Chapter 2. All the computational analysis carried out and that will be shown from this point forward was performed with resource to the **R** software. This includes the plotting of the above Figures 4.1, 4.2 and 4.3 to which corresponds the code in Appendix A.1.

4.1 A State-of-the-Art View of the Data

In this section we will apply the methodologies in Chapter 3 to our case study from a purely stationary point of view. We will not attend to the evolution in the data through the years, but only analyze the inherent structure in the extremes of this data, according to what is the current state-of-the-art. As such, all the inference will be made under this (unrealistic) assumption that the behaviour of the observations is unchanging through time. This allows us to have a broad view of the nature of the extremes of our freediving records, without the influence of external factors that might change it through the years.

Firstly, we approach the data parametrically, applying the BM, POT and LO methods from section 3.1, and following a semi-parametric analysis will be performed according to section 3.2.

According to the theoretical indications, we will denote X the random variable that represents the characteristic under study, the maximum apnea time of competitive female Static Apnea freedivers, measured in seconds. Let us assume this r.v. has an underlying distribution function F we here consider, without proof, stationary. We have $n = 795$ independent (and identically distributed) observations of the referred variable for the same number of freedivers. These are our base conditions, valid for all the inference in this section.

4.1.1 Parametric Approach

4.1.1.1 Gumbel Method

As stated before, we have $n = 795$ i.i.d. observations of the r.v. X the maximum apnea time of competitive female Static Apnea freedivers, one for each competitive freediver. This characteristic could be registered several times for each diver (as in the original rankings) but not without high correlation of those observations. Let us assume these samples for each freediver exist but we cannot access them. It is then natural to think of our observations as the sample maximum for each diver, so we can consider them the *block maximum*, if we think of each diver as its own block (not temporally defined, as it is usual in this approach).

So, we have $(y_1, \ldots, y_m)$ the observed i.i.d. sample of maxima, where $m = n = 795$ is our number of blocks (of maxima observations) and the r.v. Y is defined as $Y \equiv M_n = \max(X_1, \ldots, X_k)$ where k is the unknown sample length for each diver. We will then try fitting a GEV distribution with parameters (ξ, μ, σ) to this sample.

Preliminary Analysis Before applying the procedures of the Block Maxima approach, we will try to get a better understanding of the tail of the underlying distribution F. A basic statistical principle is to get the histogram of the sample and to analyze its approximate shape against that of known distributions (here considering the extreme values distributions that derive from the GEV distribution). It is also commonly considered good practice to plot and analyze the box-plot. These plots, seen in Figure 4.4, correspond to the code in Appendix A.2.

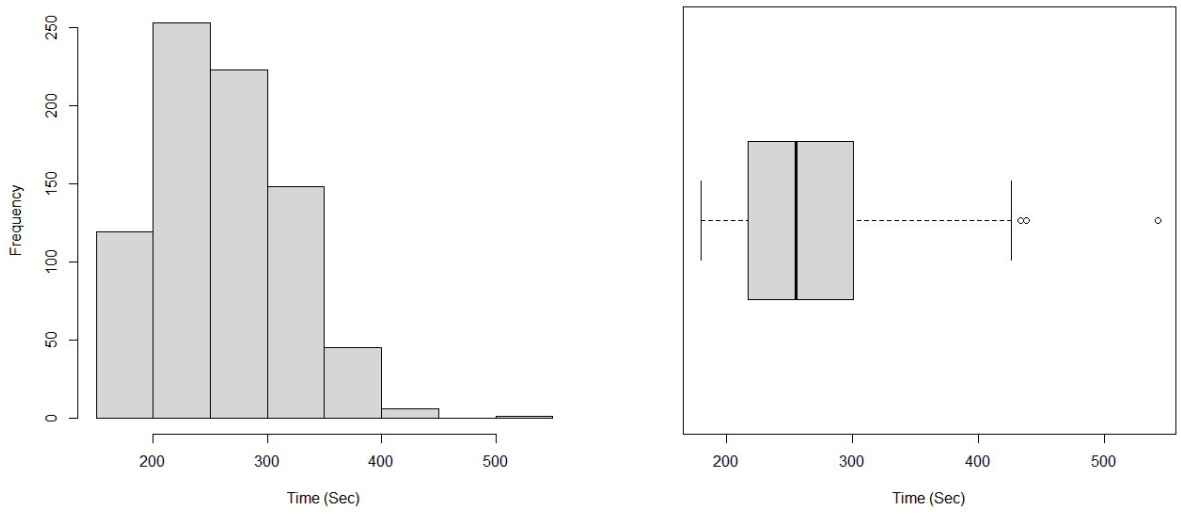

Figure 4.4: Histogram (left) and Box-plot (right) of the female SA freedivers' individual best records

It is clear in the representations of Figure 4.4 a positive asymmetry and we can see a possibly light or exponential right tail. The box-plot shows three outlier candidates and a larger concentration of observations around the 250 second mark. The histogram shows an abrupt reduction of the bars' size to the right of the 350 second mark. This suggests that a heavy tailed distribution might not be appropriate.

To get a clearer idea of the type of extreme distribution that better fits the maxima data, we resort to the qq-plots, as introduced in section 3.1.1 (details in Gomes et al. (2013a)). Firstly, following the recommendation of Beirlant et al. (2004), we use an Exponential qq-plot to identify the weight of the tail of the underlying distribution, plotting the empirical quantiles $y_{i:m}$ (sorted maxima) and the theoretical Exponential quantiles $-\log(1-p_i)$, with $p_i := {}^{i}/_{(m+1)}$ the chosen definition of the plotting positions, allied with the Mean Excess plot, obtained by plotting the curve of $(y_{m-k:m}, \hat{e}_m(y_{m-k:m}))$ for $k = 1, \ldots, m-2$ as defined in section 3.1.2.6 (the last 3 points of the ME-plot were omitted as they disrupted the visualization). This was performed with resource to the code in Appendix A.3.

As we can see in Figure 4.5, a line was fitted to the Exponential qq-plot (using the Ordinary Least Squares method of the software) that not only provides estimates of the scale and location parameters of the distribution through the slope and intercept $(\hat{\mu}, \hat{\sigma}) = (208.6084, 52.8396)$, but also helps to see the clear convex pattern of the qq-plot. This is concordant with the pattern of decreasing monotony in the ME-plot since, according to the cited authors, both the convex pattern in the qq-plot and the decreasing monotony pattern in the ME-plot suggest lighter than Exponential tails. The correlation printed in the qq-plot indicates the degree of linear correspondence between the empirical quantiles and the theoretical Exponential quantiles, which in this case is around 95%.

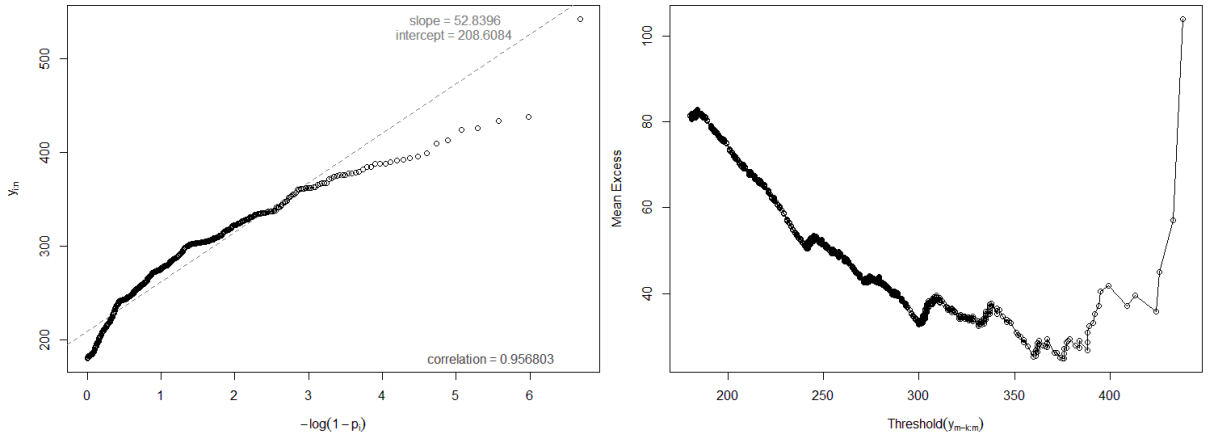

Figure 4.5: Exponential QQ-plot (left) and ME-plot (right) of the female SA freedivers' individual best records

So, we are now more confidant in discarding the possibility of an underlying heavy-tailed distribution for Y, so we are left with distributions in the Gumbel and Weibull max-domains.

To determine if the Gumbel distribution itself is an appropriate candidate for F (since the Gumbel distribution has a lighter than Exponential tail), we resort to a Gumbel qq-plot, as presented in section 3.1.1. Once again, a Least Squares line was fitted to the plot, by the code in Appendx A.4, providing the estimates $(\hat{\mu}, \hat{\sigma}) = (236.8753, 42.4374)$ for the location and scale parameters, given by the intercept and slope of the line, respectively.

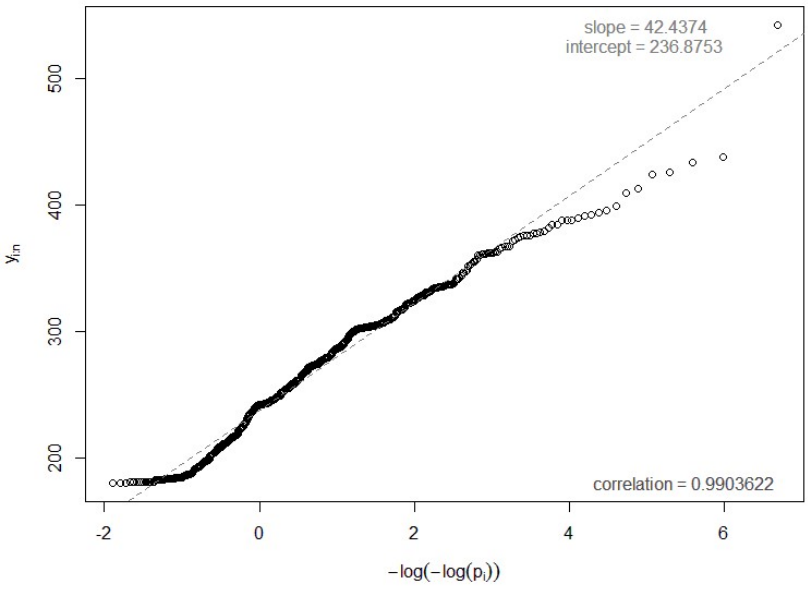

Figure 4.6: Gumbel QQ-plot of the female SA freedivers' individual best records

The correlation between the empirical and theoretical quantiles is over 99%, larger than for the Exponential fit, which indicates we are in the correct path to finding the most appropriate domain of attraction for the r.v. Y. Figure 4.6 shows that there is a pattern very close to linearity in the central quantiles, but when we look to the right tail values, there is a large deviation from linearity. There is an apparent overall convex pattern, than suggests that the tail of the underlying distribution is possibly even lighter than that of the Gumbel distribution, meaning a distribution in the Weibull max-domain might be more suitable.

This is a suggestion that the value of the Extreme Value Index could be negative, what should be taken into consideration when trying to fit a GEV distribution to the sample. A GEV qq-plot is obtained by plotting the empirical quantiles $y_{i:m}$ (sorted maxima) against the theoretical GEV quantiles $\frac{(-\log(-p_i))^{-\xi}-1}{\xi}$ where p_i are again the plotting positions. The problem is we do not know the value of the EVI ξ and as such cannot draw this plot. We need to somehow get a preliminary estimate for this parameter. Using once again the Ordinary Least Squares method, and having the EVI vary from -1 to 0.55 (since we expect a light tail, our emphasis is on the negative values of ξ) we calculate the correlation between $y_{i:m}$ and $\frac{(-\log(-p_i))^{-\xi}-1}{\xi}$, that is, the correlation on the qq-plot derived of fitting a GEV distribution to the sample for each considered value of ξ. This method, suggested in Beirlant et al. (2004), is based on the assumption that a higher correlation indicates a better fit. The code in Appendix A.5 yields the following correlation plot, where the maximum correlation and corresponding value of the EVI were marked.

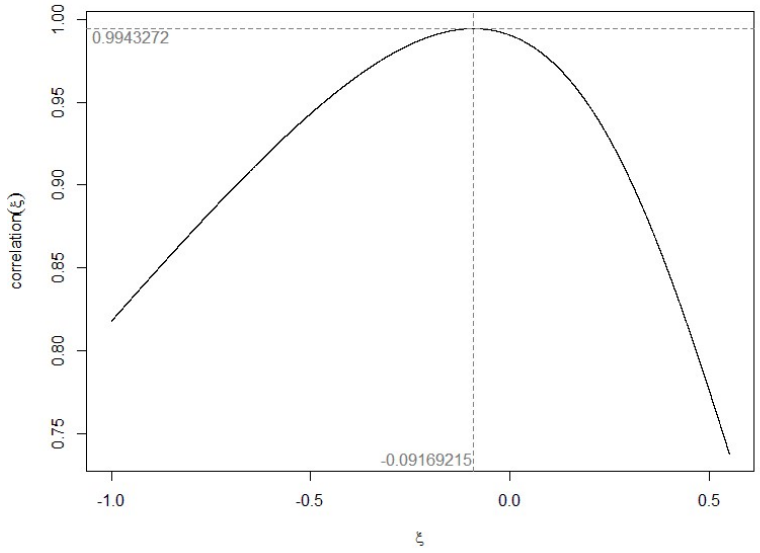

Figure 4.7: Correlation plot for the GEV family QQ-pot of the female SA freedivers' individual best records

Considering then the optimal value of the EVI as $\hat{\xi} = -0.09169215$ in the sense that it yields the largest (high) correlation of approximately 99.4% for the fit, we have the corresponding qq-plot, which shows a much more extended linear pattern indicating a better fit of the GEV distribution with that EVI to Y (corresponding code in Appendix A.6). There is a significant

deviation from linearity corresponding to the sample maximum due to the much larger magnitude of that value comparing to all the others. This point corresponds to the personal best of world recordist Natalia Molchanova.

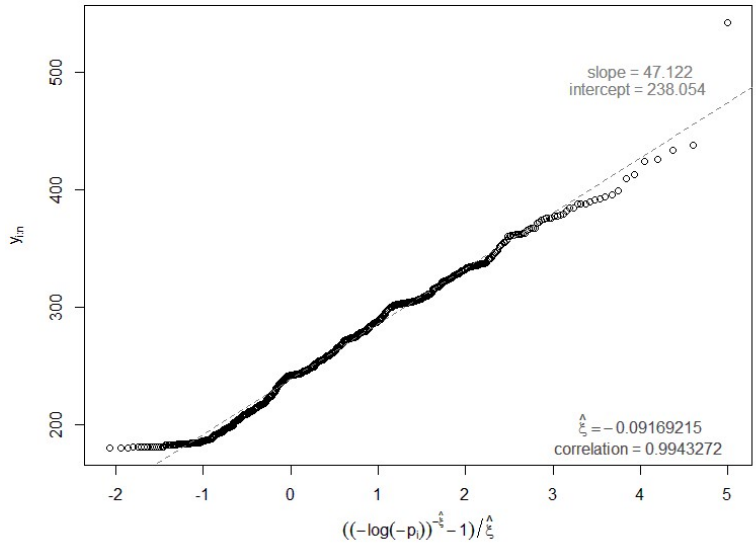

Figure 4.8: GEV QQ-plot for $\hat{\xi} = -0.09169215$ of the female SA freedivers' individual best records

The Least Squares fit of the quantile points in this plot suggests the estimates $(\hat{\mu}, \hat{\sigma}) = (238.054, 47.122)$ for the location and scale parameters of the considered underlying GEV distribution, with $\hat{\xi} = -0.09169215$. Note that, as we suspected, the optimum value of the EVI given by the correlation plot is negative, although fairly close to 0, which allows us to completely discard the Fréchet domain of attraction but not to conclude with certainty whether F in on the Weibull or Gumbel max-domain.

This preliminary analysis was not very conclusive, although enough for a Fréchet-type distribution to be out of the question and to infer that we are dealing with lighter right tailed distributions. Thus, we will consider as our preliminary estimates the $(\hat{\mu}, \hat{\sigma})$ corresponding to the Gumbel and GEV with $\hat{\xi} = -0.09169215$ fits presented. The corresponding probability density functions of this estimated distributions were laid over the histogram of the data in Figure 4.9 for a better visualization of the quality of the adjustments – see Appendix A.7.

Both in the Gumbel and in the GEV qq-plots (Figures 4.6 and 4.8 resp.), there is an initial "flatness" of the first quantiles that doesn't quite adjust to the fitted straight line in either case. This is provoked by a high concentration of equal values in the lower fraction of the sample – many of the more average freedivers have the same best record, equal to or not much larger than our competitive starting level of 180 seconds. This is not concerning.

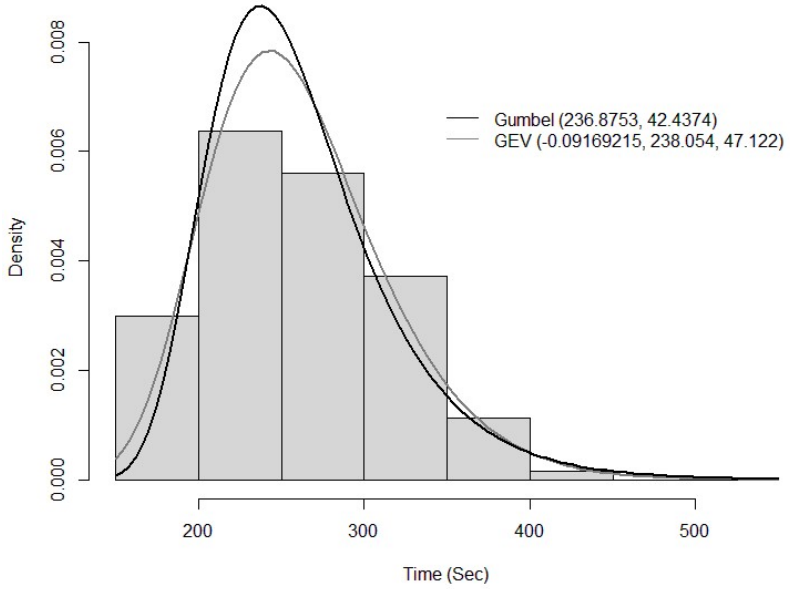

Figure 4.9: Histogram with Gumbel and GEV fitted p.d.f. for the female SA freedivers' individual best records

Statistical Choice of Domain of Attraction Since we were unable in the preliminary analysis to undoubtedly choose a domain of attraction for the underlying distribution F, we must perform tests of statistical choice to decide between the two extreme types of distributions that are still in play: the Gumbel domain of attraction, corresponding to $\xi = 0$, and the Weibull domain, corresponding to $\xi < 0$.

Assuming an underlying GEV-type distribution to the r.v. Y for some real ξ, we will apply the testing procedures described in section 3.1.1.5 for the following sets of hypothesis:

$$H_0 : \xi = 0 \qquad versus \qquad H_1^{(1)} : \xi \neq 0 \tag{4.1}$$

and

$$H_0 : \xi = 0 \qquad versus \qquad H_1^{(2)} : \xi < 0 \, . \tag{4.2}$$

Some of the tests introduced, such as the Likelihood Ratio Test, require the prior Maximum Likelihood estimation of the GEV distribution's parameters. For the time being, that estimation is performed with resource to the functions in the **R** package *fitdistrplus*, firstly assuming an underlying distribution to the data from the Gumbel family, and then considering a distribution from the unrestricted GEV family. The obtained estimates, which are quite similar in both cases, are compiled in Table 4.2. All the code referring to this section can be found in Appendix A.8.

Let us first regard the tests for the two-sided alternative (4.1), at the usual significance levels of 5% or 1%.

Table 4.2: ML estimates from the *fitdistrplus* package for the location, scale and shape parameters for the fitted Gumbel and GEV distributions to the female SA freedivers' individual best records.

	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
Gumbel	235.65797	44.85943	-
GEV	237.2037489	45.87483486	-0.06279132

Table 4.3: Test results for the hypothesis in (4.1) for the female SA freedivers' individual best records.

Test	Observed Statistic	Observed p-value	Decision
LR	4.266774	0.03886465	Rejection of H_0 at a 5% level; Non-rejection of H_0 at a 1% level
Rao's Score	2.388051	0.122266	Non-rejection of H_0 at a 5% level
LAN	-1.546425	0.122002	Non-rejection of H_0 at a 5% level

Both the Rao's Score test and the Local Asymptotically Normal test point towards the non-rejection of the Gumbel hypothesis when confronted with a two-sided non-Gumbel alternative, at the comfortable asymptotic level 5%, although we can consider this a borderline decision if working at a 10% asymptotic significance level. The Likelihood Ratio test is more conservative, and only points to the non-rejection of the Gumbel hypothesis versus the $H_1^{(1)}$ alternative at levels inferior to 3%, making it once again a borderline decision.

It seems to be the suggested conclusion that we should not reject the Gumbel hypothesis in favour of a non-Gumbel GEV distribution, but since the decisions aren't clear, depending on the significance level chosen from the usual 10%, 5% and 1%, we resort to the test based on the Gumbel Statistic as presented in Gomes and Fraga Alves (1996). For a better understanding of the behaviour of this statistic, which depends on the sample fraction used for its computation, we plotted in Figure 4.10 its values against the number of top o.s.'s used, and overlaid the exact and asymptotical critical points given in the referred paper.

Contrary to what we hoped, this isn't much clearer than the previous tests, since beyond a sample fraction of almost half the number observations, that is, almost 400 top o.s.'s used in the statistic's computations, we begin to reject the Gumbel hypothesis in favour of a non-Gumbel underlying distribution F. Between the use of around 50 and 380 o.s.'s, the decision is of non-rejection of the Gumbel null hypothesis. Since it is unclear what is the ideal sample fraction we should consider, we cannot draw a definite conclusion from this statistic either.

Let us see how the tests perform for the one-sided Weibull domain alternative hypothesis in 4.2, for the same significance levels as usual.

As we can observe in Table 4.4, once again both the Rao's Score and the LAN tests point towards the non-rejection of the Gumbel hypothesis at a 5% significance level, even now versus a Weibull domain alternative. Still, this decision can be one more time considered as borderline,

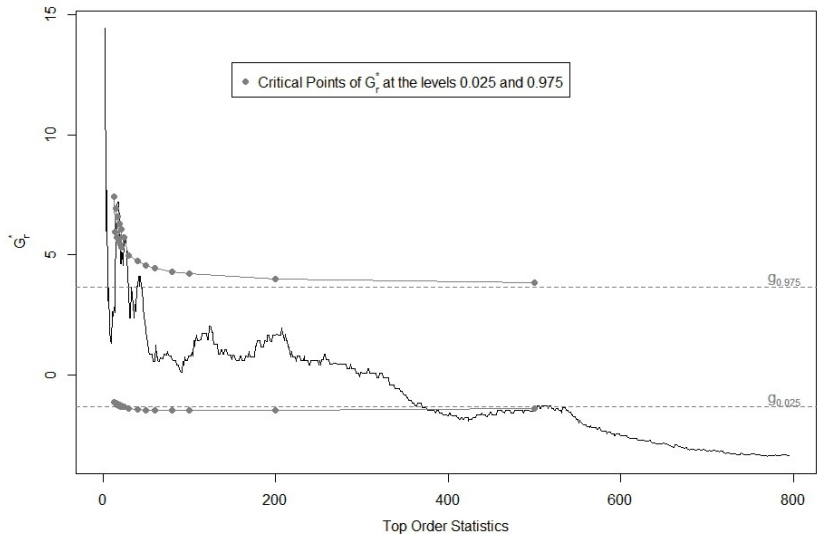

Figure 4.10: Gumbel test statistic and corresponding two-sided critical points for the female SA freedivers' individual best records

since for significance levels over approximately 6.1% the decision if favorable to a Weibull domain distribution underlying the r.v. Y. Only the test based on the Gumbel statistic (as presented by Tiago de Oliveira and Gomes (1984)) undoubtedly decides on the non-rejection of a Gumbel distribution at any usual level of significance.

Table 4.4: Test results for the hypothesis in (4.2) for the female SA freedivers' individual best records.

Test	Observed Statistic	Observed p-value	Decision
Rao's Score	-1.545332	0.061133	Non-rejection of H_0 at a 5% level
LAN	-1.546425	0.06100099	Non-rejection of H_0 at a 5% level
Gumbel Statistic(*)	1.975694	0.8705196	Non-rejection of H_0 at a 5% level

(*) as in Tiago de Oliveira and Gomes (1984)

If we plot now the observed values of the Gumbel Statistic as presented in Gomes and Fraga Alves (1996), with the overlay of the critical points for the one-sided rejection region, as can be seen in Figure 4.11, we are anew faced with a conflicting conclusion: beyond a sample fraction of almost half the number observations, that is, almost 400 top o.s.'s used in the statistic's computations, we begin to reject the Gumbel hypothesis in favour of a Weibull-type underlying distribution F. For the statistics calculated with less than that number of o.s.'s, the decision is of non-rejection of the Gumbel null hypothesis. As such, it is once more unclear which conclusion we should draw from this testing procedure.

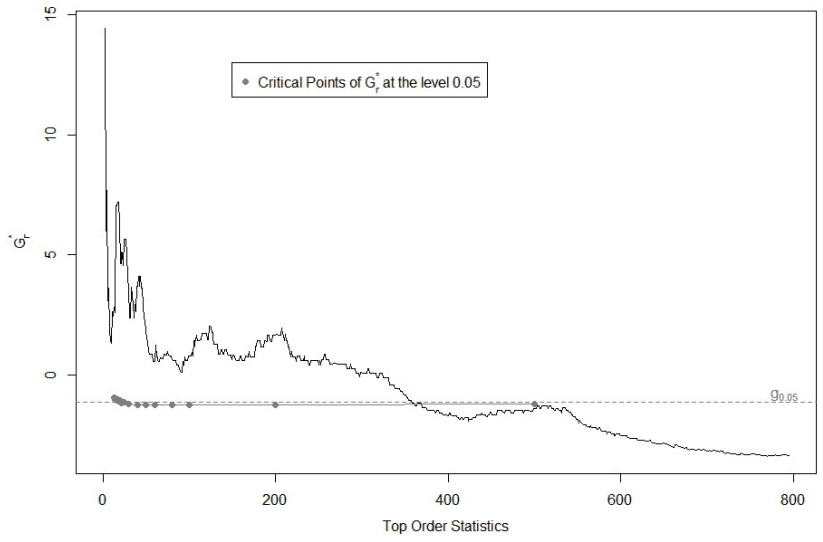

Figure 4.11: Gumbel test statistic and corresponding one-sided critical points for the female SA freedivers' individual best records

We are then left with we the possibility of performing the goodness-of-fit tests for the Gumbel distribution as introduced in section 3.1.1.5. As before, the corresponding code can be found in Appendix A.8. The results of this tests are condensed in Table 4.5.

Table 4.5: Goodness-of-fit of the Gumbel distribution test results for the female SA freedivers' individual best records.

Test	Observed Modified Statistic	Critical Point 5%	Critical Point 1%	Decision
Kolmogorov-Smirnov	1.347547	0.874	1.007	Rejection of the Gumbel model
Cramér-von Mises	0.4604857	0.124	0.175	Rejection of the Gumbel model
Anderson-Darling	3.64682	0.757	1.038	Rejection of the Gumbel model

The goodness-of-fit tests for the Gumbel model give us contradictory information when compared to the previous hypothesis tests performed. The decision of these testing procedures is without question the rejection of a good fit of the Gumbel distribution to the data from the r.v. Y, at every usual significance level not inferior to 1%.

We suggest that this problem of contradiction comes from the fact that our sample is too large, having 795 maxima observations. As referred in the literature (for instance, see Razali and Wah (2011)), goodness-of-fit tests such as the ones here performed, that are based on the empirical d.f., do not perform very well with large samples, since they become more sensitive to

even very small deviations from the hypothesised distribution. When the sample size increases, so does the power of the test (that is, the probability of not committing a type II error – the probability of rejecting the null hypothesis when the alternative is true), leading to the easy rejection of the null hypothesis.

Having under advisement all the testing procedures applied in this section, we decided it unwise to dismiss either domain of attraction (Gumbel or Weibull), being that further testing would be necessary for drawing more definitive conclusions. This decision prevents us from concluding if our variable has a finite or infinite right endpoint, but we can here make use of a common sense argument: given that it is physiologically unlikely (impossible, to our knowledge) for a human being to hold their breath indefinitely, we can safely assume the maximum apnea time of any freediver to be finite, and as such its underlying distribution should have a finite right endpoint.

Estimation of Parameters and Other Extreme Value Indicators For inference purposes, given the inconclusiveness of the testing procedures used in the previous section, we will consider both possibilities in play so far: an attraction of the underlying d.f. F to the Gumbel max-domain, corresponding to $\xi = 0$, and an alternative attraction to the Weibull max-domain, corresponding to $\xi < 0$. Although, as stated before, we should keep in mind that we are dealing with a variable that should biologically have a finite right endpoint.

But before we begin the inference about the female competitive SA freedivers' individual best records, we must decide which indicators are in our interest to estimate, obviously beyond the basic estimation of the location and scale parameters, as well as the tail index.

Having that the maximum of the Y_i sample corresponds to the world record of 542 seconds – Natalia Molchanova's 2013 9 minutes and 2 seconds breath hold – it would be interesting knowing, given the current state-of-the-art, the probability that this value will be surpassed (i.e. that a female freediver will set her personal best record above this time). So we want to know the **exceedance probability of 542** – $P[Y > 542]$.

Another possibly interesting information is the values that are exceeded with a very small probability – the extremal quantiles. Associated with these quantities, we can have the return levels, which are basically just another interpretation of the same statistical entity. Let us apply both interpretations, when we estimate the individual female best competitive Static Apnea time that is registered with a 0.01% probability, that is the **extremal quantile of probability** $(1 - 0.0001)$ – $\chi_{0.0001}$; and when we estimate the apnea individual record time that is exceeded, on average, once every 100 best personal records, that is the **return level of** 100 **personal records** – $U(100)$. This return level interpretation reflects the fact that our "blocks" correspond to freedivers, and are not defined by any time period, so the way we read this information is a bit more far-fetched, but still interesting. Recall that the return level $U(100)$ corresponds to the extremal quantile $\chi_{0.01}$.

Finally, the last parameter we are here very interested in estimating is the absolute maximum apnea personal record time that could possibly be set by a SA competitive female freediver. This is statistically represented by the right endpoint, which should be finite when the case is that

the EVI's estimate from the GEV fitted model is negative (and we expect it to happen for every estimation method employed). As such, we will also estimate the **finite right endpoint** – x^F.

For each model fitted (Gumbel or GEV) we will use 3 estimation methods: the first is based on the preliminary estimates from the qq-plots shown in the previous section, and also both methods mentioned in sections 3.1.1.1 and 3.1.1.2 – the Maximum Likelihood estimation and the Probability Weighted Moments estimation, respectively. Specifically for the ML method, we made use of pre-programmed fitting functions from 4 different packages in the **R** software: *fitdistrplus* (already employed in the previous section on the statistical testing for the max-domain), *fExtremes*, *ismev* and *evd*. The objective was verifying that the specific programming of each function did not interfere significantly with the estimation results.

Let us start by fitting the Gumbel model to our sample of maxima, using the expressions related to $\xi = 0$ given in section 3.1.1. All the **R** code used for the fitting and which produced the following estimates and plots can be found in A.9. Table 4.6 comprises the obtained estimates for the Gumbel model parameters $(\hat{\mu}, \hat{\sigma})$ and for the mentioned indicators of interest $\widehat{P[Y > 542]}$, $\widehat{\chi_{0.0001}}$ and $\widehat{U(100)}$ through the preliminary, ML and PWM estimation methods.

Table 4.6: Estimates for the Gumbel fit to the female SA freedivers' individual best records.

Method	Package	$\hat{\mu}$	$\hat{\sigma}$	$\widehat{P[Y > 542]}$	$\widehat{\chi_{0.0001}}$	$\widehat{U(100)}$
Preliminary	-	236.8753	42.4374	0.0007538085	627.7361	432.0937
ML	*fitdistrplus*	235.65797	44.85943	0.001081427	648.8263	442.018
	fExtremes	235.69408	44.87459	0.001084796	649.0021	442.1239
	ismev	235.65959	44.86871	0.001082994	648.9135	442.0624
	evd	235.65835	44.86011	0.001081548	648.833	442.0215
PWM	-	235.8835	43.91458	0.0009385304	640.3495	437.8971

Analysing Table 4.6 above we can immediately note that the ML parameter estimation performed by the four used packages is concordant, and differences occur only beyond the second decimal place, so they are not significant.

More importantly, we can see that all three estimation methods yield very similar estimates for the intrinsic parameters of location and scale for the Gumbel model – placing the location at around the 235 seconds and the scale at around 43. This shows us that the preliminary approach, based on the qq-plots, can in fact be an effective and even reliable tool, to some extent, for assessing extreme value data, giving fairly acceptable estimates for the parameters. The most similar estimates are the ones from the ML and PWM methodologies: the estimates of μ differ only on approx. 2 decimal points and those of σ differ just a little more significantly, on approximately a unite (one second).

However, even the small differences observed between the model parameter's estimates for each estimation method can translate into more relevant disparities in the derived estimates for the indicators of interest. Take, for example, the estimation of the extremal quantile $\widehat{\chi_{0.0001}}$. There is a gap of almost 20 seconds between the preliminary estimation of roughly 627 seconds and the ML estimates of around 649 seconds. It may not seem much of a difference, but when it

comes to breath holding and especially after 600 seconds of not breathing, anyone would agree that 20 seconds represent a significant amount of apnea time. Also, despite the ML and PWM estimates of μ and σ being fairly close, the respective estimates for the $\chi_{0.0001}$ and $U(100)$ are more deviant, coming up to 9 and 5 seconds differences in each case.

To comment on the quality of the Gumbel adjustment we can analyse the graphical diagnostic tools that the software presents us with. We chose here to present the output of said diagnostic tools from the *fitdistrplus* and *evd* packages, for they present complementing information, represented in Figures 4.12 and 4.13, respectively. Note that this evaluation regards the fit of the distribution with its parameters estimated by the Maximum Likelihood method, but since they are similar to the ones from the other methods, we can generalize the analysis for the fit in itself, not depending on the estimation procedure.

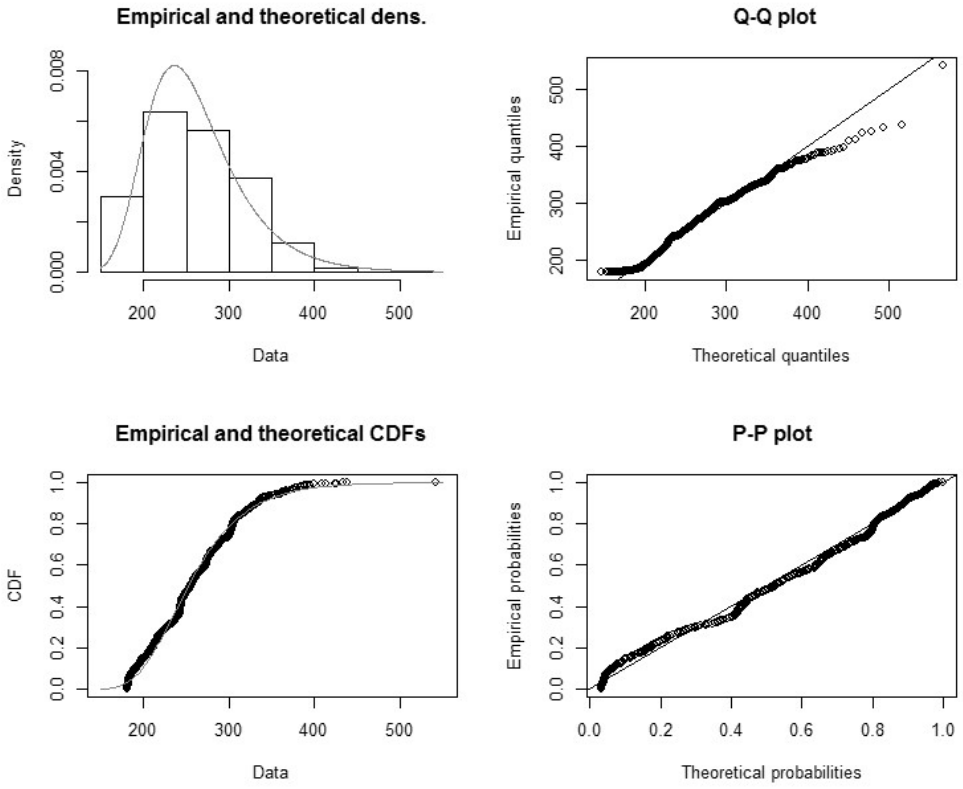

Figure 4.12: Diagnostic plots given by the *fitdistrplus* package for the Gumbel fit to the female SA freedivers' individual best records

These diagnostic plots show that the adjustment appears to be quite satisfactory. We can see the p.d.f. of the fitted Gumbel model is fairly approximated both to the shape of the data histogram (top-left plot of Figure 4.12) and to the empirical density function (bottom-left plot of Figure 4.13). The qq-plot is included in the output from both packages because of its usefulness in evaluating the adjustment, but focussing on the quantile plot from the *evd* package (top-right plot of Figure 4.13), we find that all the points seem to be approximately contained between the confidence bands plotted by the software, with the exception of beginning points on the left side of the plot and also some of the largest observations to the right.

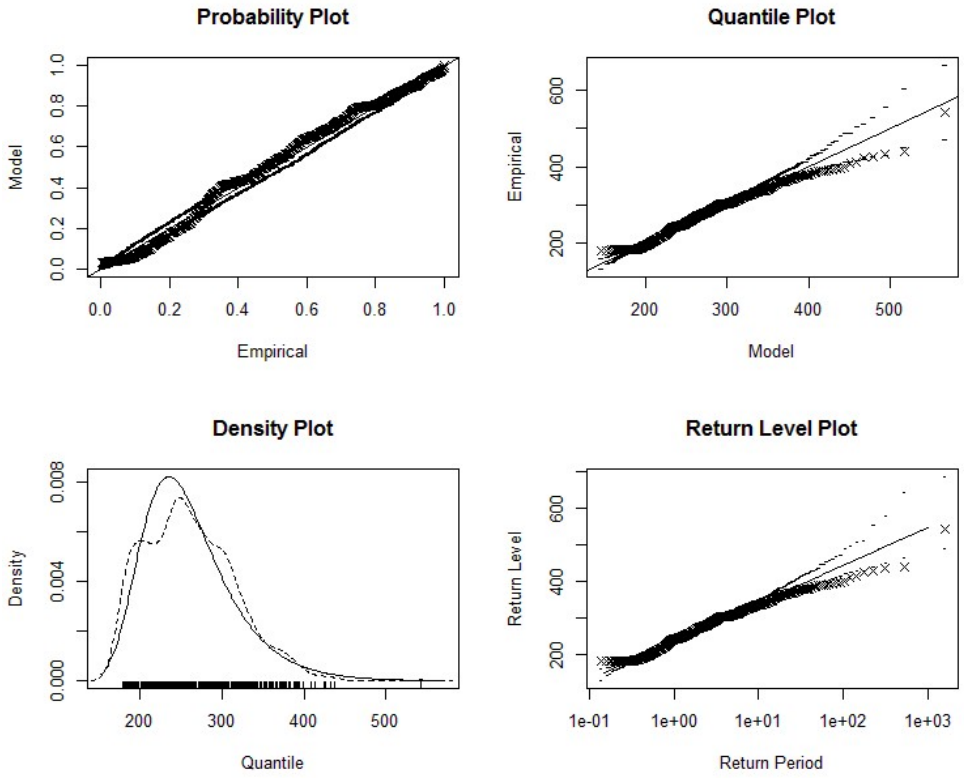

Figure 4.13: Diagnostic plots given by the *evd* package for the Gumbel fit to the female SA freedivers' individual best records

Moreover, we can see in the bottom-left plot of Figure 4.12 that the empirical and theoretical d.f.'s are reasonably concordant. We then conclude that the Gumbel distribution with the estimated parameters shown before is a quite suitable fit for the freediving maxima data.

Following the Profile Log-Likelihood procedure presented in section 3.1.1.4, we present here 95% Confidence Intervals for the parameters of this fitted Gumbel model. These were obtained with the help of the software (as exposed in Appendix A.9), which also allows the plotting of the respective profile log-likelihood functions, seen here in Figure 4.14.

The resulting 95% CI's based on the profile log-likelihood for the location and scale parameters under the Gumbel fit to the data are then

$$CI_{\mu}^{95\%}(G_0) = [232.39113\,;\, 238.97571]$$

and

$$CI_{\sigma}^{95\%}(G_0) = [42.51197\,;\, 47.40982]\,.$$

The same procedure was employed for calculating the 95% confidence intervals for the extremal quantile $\chi_{0.0001}$ and for the return level $U(100)$, based on their functional expressions written in (3.11) and (3.10) resp., in order to the model parameters μ and σ.

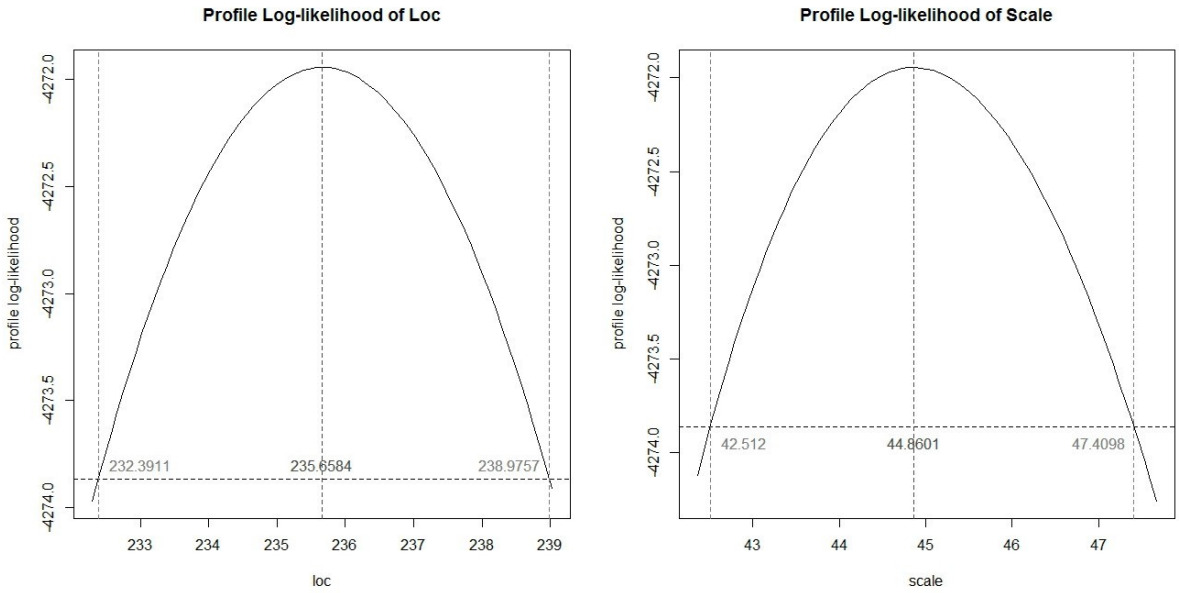

Figure 4.14: Profile Log-Likelihood plots and 95% CI's for location (left) and scale (right) parameters of the Gumbel fit to the female SA freedivers' individual best records

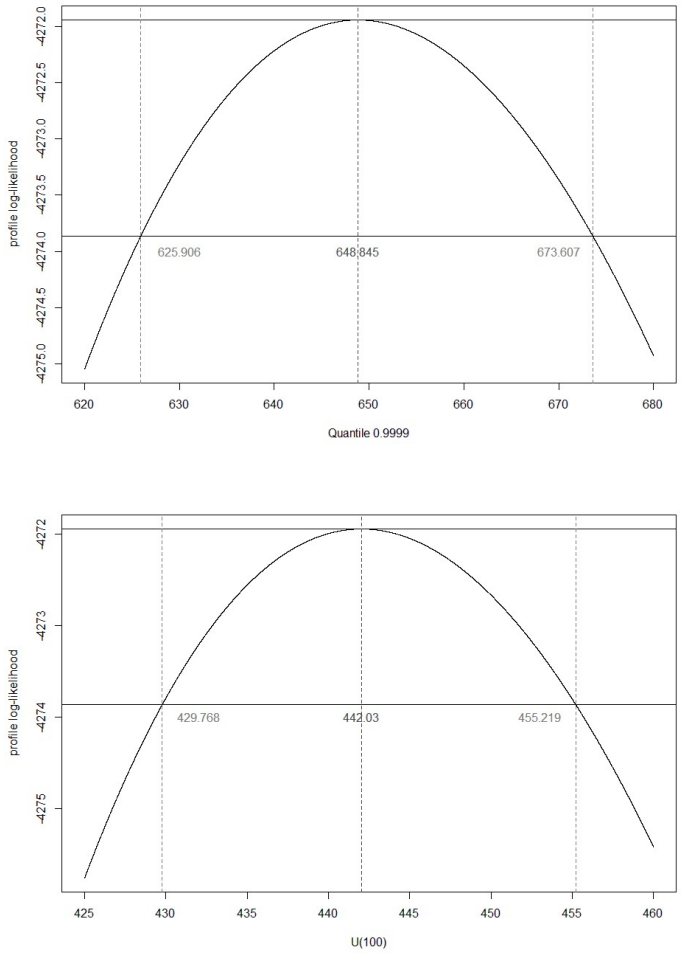

Figure 4.15: Profile Log-Likelihood plot and 95% CI for $\chi_{0.0001}$ (top) and $U(100)$ (bottom) under the Gumbel fit to the female SA freedivers' individual best records

The plots of the profile log-likelihood functions for these quantities, which can be seen in Figure 4.15, were not based in any pre-existing **R** package function, since one could not be found that performed satisfactorily – the corresponding code is once again included in Appendix A.9. The 95% CI's yielded for $\chi_{0.0001}$ and $U(100)$ under the Gumbel fit to the data are then

$$CI^{95\%}_{\chi_{0.0001}}(G_0) = [625.906\,;\,673.607]$$

and

$$CI^{95\%}_{U(100)}(G_0) = [429.768\,;\,455.219]\,.$$

In summary, according to this Gumbel fit to the maxima sample of the personal records of female SA freedivers, and in the current stationarity setup, we estimate that:

- The probability that a female freediver will set her best personal SA record above the current world record of 9 minutes and 2 seconds (542 seconds) is approximately 0.1%, which is very reduced;
- There is approximately a 0.01% probability that that a female freediver will set her best personal SA record above $\approx$ 10 minutes and 20 seconds (640 seconds), meaning this is a very unlikely mark;
- In average, about 100 female SA freedivers must set their best mark ever so that a best individual record above $\approx$ 7 minutes and 20 seconds (440 seconds) is observed.

As referred above, the GEV distribution with a possibly negative EVI is another candidate to be fitted to our maxima sample. We are guided again by the methodology in section 3.1.1, but now focussing on the presented expressions regarding the case $\xi \neq 0$, that is, discarding the Gumbel domain case, already addressed. Once more resorting to preliminary, ML and PWM estimation, we comprised in Table 4.7 the obtained estimates for the GEV model shape, location and scale parameters $(\hat{\xi}, \hat{\mu}, \hat{\sigma})$ and for the mentioned indicators $\widehat{P[Y > 542]}$, $\widehat{\chi_{0.0001}}$, $\widehat{U(100)}$ and, if applicable (that is, for the cases when $\hat{\xi} < 0$) also $\widehat{x^F}$. The code developed in the **R** software for this GEV fitting and all the subsequent analysis and plots can be found in Appendix A.10.

Table 4.7: Estimates for the GEV fit to the female SA freedivers' individual best records.

Method	Package	$\hat{\xi}$	$\hat{\mu}$	$\hat{\sigma}$	$\widehat{P[Y > 542]}$	$\widehat{\chi_{0.0001}}$	$\widehat{U(100)}$	$\widehat{x^F}$
Prelim.	-	-0.09169215	238.05383	47.12241	5.76061e-05	531.1077	414.9107	751.9736
ML	*fitdistrplus*	-0.06279132	237.20374890	45.87483486	0.0001844114	558.0533	420.491	967.7957
	fExtremes	-0.06275239	237.20621763	45.87758286	0.0001848751	558.127	420.5201	968.2952
	ismev	-0.06269218	237.19388858	45.87829372	0.0001853191	558.2001	420.5348	968.9963
	evd	-0.06287844	237.21759969	45.88067286	0.0001840578	557.9916	420.4932	966.8902
PWM	-	-0.1061981	238.1244	48.00463	2.742784e-05	520.1818	412.8203	690.1533

We find once again that the ML parameter estimation performed by the four used packages is concordant, as expected.

The first aspect that comes to attention when analyzing Table 4.7 is that all estimates of the tail index ξ are negative, confirming our belief that a short tailed model would be the most suitable from the GEV family. However, said estimates are relatively close to 0, which means the fitted GEV distribution will be close to the Gumbel model.

Comparatively to what happened with the results of the Gumbel fit, in this case there appear to be more significant differences between the ML and PWN estimates for the core parameters of the model, even more so than the ones found when comparing the preliminary and ML estimates. The ML method provides with the largest estimate for the EVI (the less negative one), approximately 3 hundredths higher than the preliminary ξ estimate (obtained from the correlation plot in Figure 4.7) and 4 hundredths higher than the PWM ξ estimate. The PWM and preliminary estimates for the location parameter μ are very similar, both differing from the ML corresponding estimate in almost a unit (one second). The most evident differences are regarding the scale parameter, whose ML estimates are around 1.3 and 2.2 seconds smaller than the preliminary and PWM ones.

These differences imply large disparities in the estimation of the other indicators of interest, especially in those for the finite right endpoint of the distribution. Furthermore, the estimates of the location and scale parameter are reasonably similar to the ones obtained for the Gumbel fit, but sufficiently different that they yield estimates, for example, for the exceedance probability of the sample maximum completely discordant.

Regarding $\widehat{\chi_{0.0001}}$, the ML and PWM differ in more than 40 seconds. The argument given when analysing the Gumbel estimates is once again valid here for declaring this as a significant difference. For the $\widehat{U(100)}$, the difference is more subtle but even so representing an additional 8 seconds of breath hold according to the ML method in comparison to the PWM estimate. The ML method also produces the largest $\widehat{P[Y > 542]}$, even so being approx. 100 times smaller than the corresponding estimate for the Gumbel fit.

In what concerns the right endpoint estimate, the three methods are very evidently disagreeing. According to the ML method, it is expected that the highest personal record a female SA freediver can possibly achieve is 968 seconds, that is, 16 minutes and 8 seconds. The preliminary methodology yields a more conservative estimate of 751 seconds (12 minutes and 31 seconds) and the PWM gives us an even lower right endpoint estimate of 690 seconds (11 minutes and 30 seconds). It is therefore clear that an almost 5 minute difference is significant, given the nature of the variable it refers to. Given that the ML estimate represents the largest possible limit, we will consider it our best case scenario estimate.

Analogously to what was done for the Gumbel fit, we will now analyze the diagnostic plots provided by the software, using the same previously mentioned packages and once again referring to the results of the ML estimation. Some of the interest lays here in comparing the quality of the GEV adjustment as shown by Figures 4.16 and 4.17 to the already assessed quality of the Gumbel adjustment as shown by Figures 4.12 and 4.13.

A very indistinguishable plot from its Gumbel counterpart is the overlap between the empirical and theoretical distribution functions on the bottom-left plot of Figure 4.16, since the approximation appears to be just as good as the one in Figure 4.12.

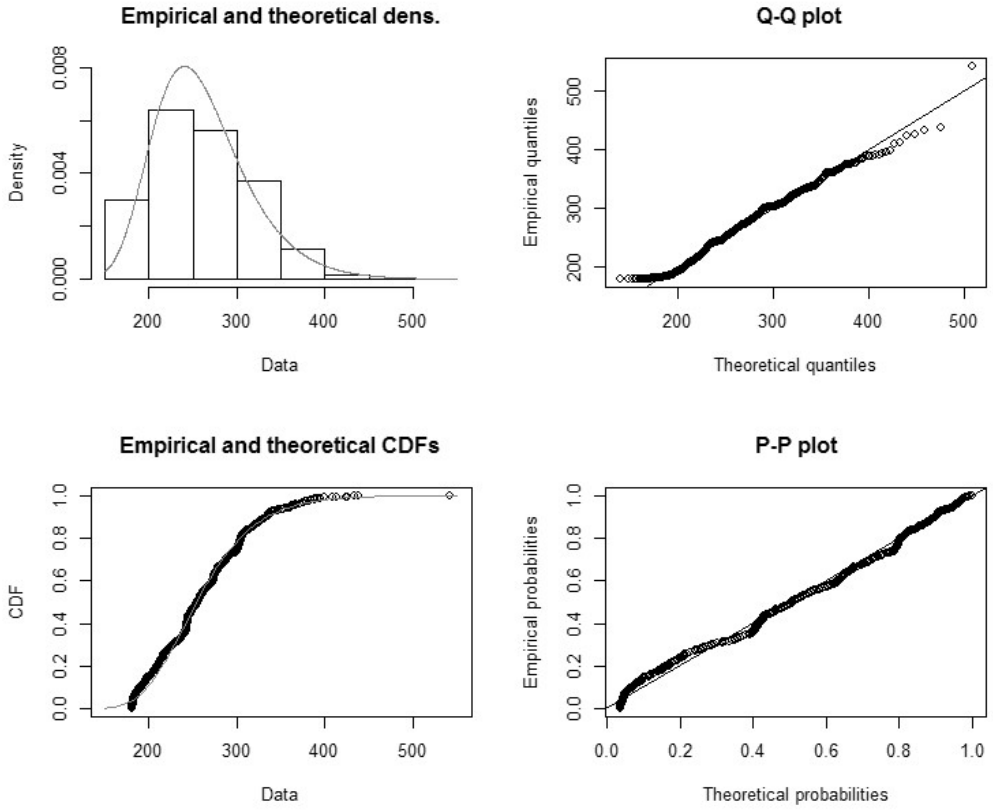

Figure 4.16: Diagnostic plots given by the *fitdistrplus* package for the GEV fit to the female SA freedivers' individual best records

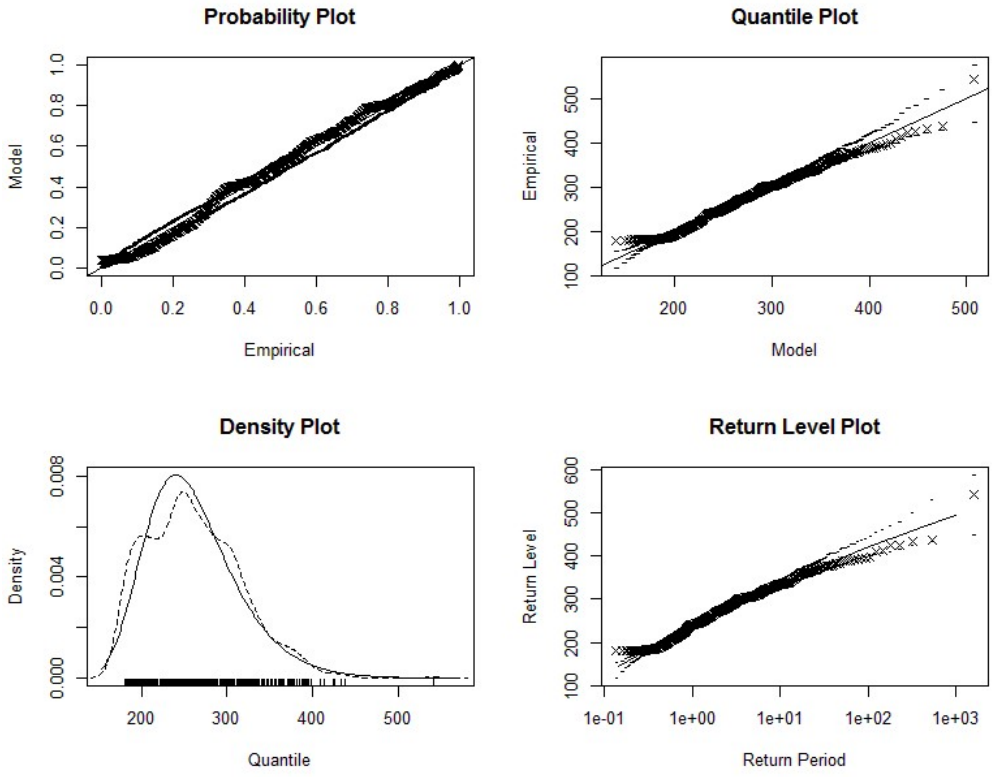

Figure 4.17: Diagnostic plots given by the *evd* package for the GEV fit to the female SA freedivers' individual best records

The fit of the adjusted density function to the histogram of the maxima data (top-left plot of Figure 4.16) and to the empirical p.d.f. (bottom-left plot of Figure 4.17) is reasonable and at first glance appears to be again indistinguishable from the adjustment in the corresponding plots for the Gumbel fit. This is not surprising, since we have seen that the estimates of the parameters for both distributions are not very different and even the preliminarily adjusted density curves in Figure 4.9 are very close. So this information is not enough for concluding on the better adjusted distribution.

However, analyzing the probability and quantiles plots we see that the GEV fit is indeed a bit better than the Gumbel adjustment. Take, for instance, the bottom-right pp-plot in Figure 4.16, where there seems to be a better linearity between the theoretical and empirical probabilities than in the same plot in Figure 4.12 for the Gumbel adjustment. The superior quality of the GEV adjustment is the most evident in the qq-plot on the top-right of Figure 4.17 and especially in the right end portion of the plotted points, more clearly contained in the confidence bands than in the counterpart plot in Figure 4.13.

So the conclusion here seems to point in the direction of a better adjusted GEV fit (from the Weibull max-domain) when compared to the Gumbel alternative fit.

Following once more the profile log-likelihood procedure already mentioned before, we computed 95% confidence intervals for the three core parameters (ξ, μ, σ) of the GEV distribution. These were once more obtained with the help of the software (as exposed in Appendix A.10), as were the plots of the respective profile log-likelihood functions, seen here in Figure 4.18. Moreover, and as before, CI's for the indicators $\chi_{0.0001}$ and $U(100)$ were also constructed, and plotted the corresponding profile log-likelihood functions in Figure 4.19, but this time with the help of the *fExtremes* package, which includes a function that performs correctly for the GEV fit, which did not happen for the Gumbel case – details in Appendix A.10. Considering then the ML estimates for the intrinsic parameters (ξ, μ, σ), the 95% CI's based on the profile log-likelihood under the GEV fit to the data are

$$CI_{\xi}^{95\%}(G_{\hat{\xi}}) = [-0.1117133\,;\, -0.0036494]\,,$$

$$CI_{\mu}^{95\%}(G_{\hat{\xi}}) = [233.6072952\,;\, 240.8575518]\,,$$

$$CI_{\sigma}^{95\%}(G_{\hat{\xi}}) = [43.3133923\,;\, 48.6395775]\,,$$

$$CI_{\chi_{0.0001}}^{95\%}(G_{\hat{\xi}}) = [508.204\,;\, 644.6507]\,,$$

$$CI_{U(100)}^{95\%}(G_{\hat{\xi}}) = [407.1015\,;\, 450.844]\,.$$

Note that the 95% CI for the EVI does not include (even that barely) the value 0, which means the Gumbel distribution is not here included in the possible distributions for the adjustment with 95% confidence. Also note that the referred interval contains both the preliminary and PWM estimates for the shape parameter. The complete coverage of only negative values in this CI make us comfortable in believing the Weibull domain is probably the most suitable domain of attraction for our data's distribution. Both CI's for the location and scale parameters are positioned more to the right when compared with the respective intervals $CI_{\mu}^{95\%}(G_0)$ and $CI_{\sigma}^{95\%}(G_0)$. Although

having higher values for interval limits, they respectively include the ML estimates for μ and σ under the Gumbel assumption.

Regarding the $\chi_{0.0001}$ and $U(100)$ confidence intervals, it is first clear in Figure 4.19 that the profile log-likelihood-based intervals are not necessarily centered around the ML estimation of the corresponding parameter. The lower and upper limits of both CI's are positioned more to the left than the corresponding limits for the Gumbel fit (have lower values), so much that in the case of $CI^{95\%}_{\chi_{0.0001}}(G_{\hat{\xi}})$, the Gumbel ML estimate for $\chi_{0.0001}$ is not even included in this interval. The same is not true for the estimation of $U(100)$, showing that the further we position ourselves on distribution's tail, the more the differences between the adjustments are relevant.

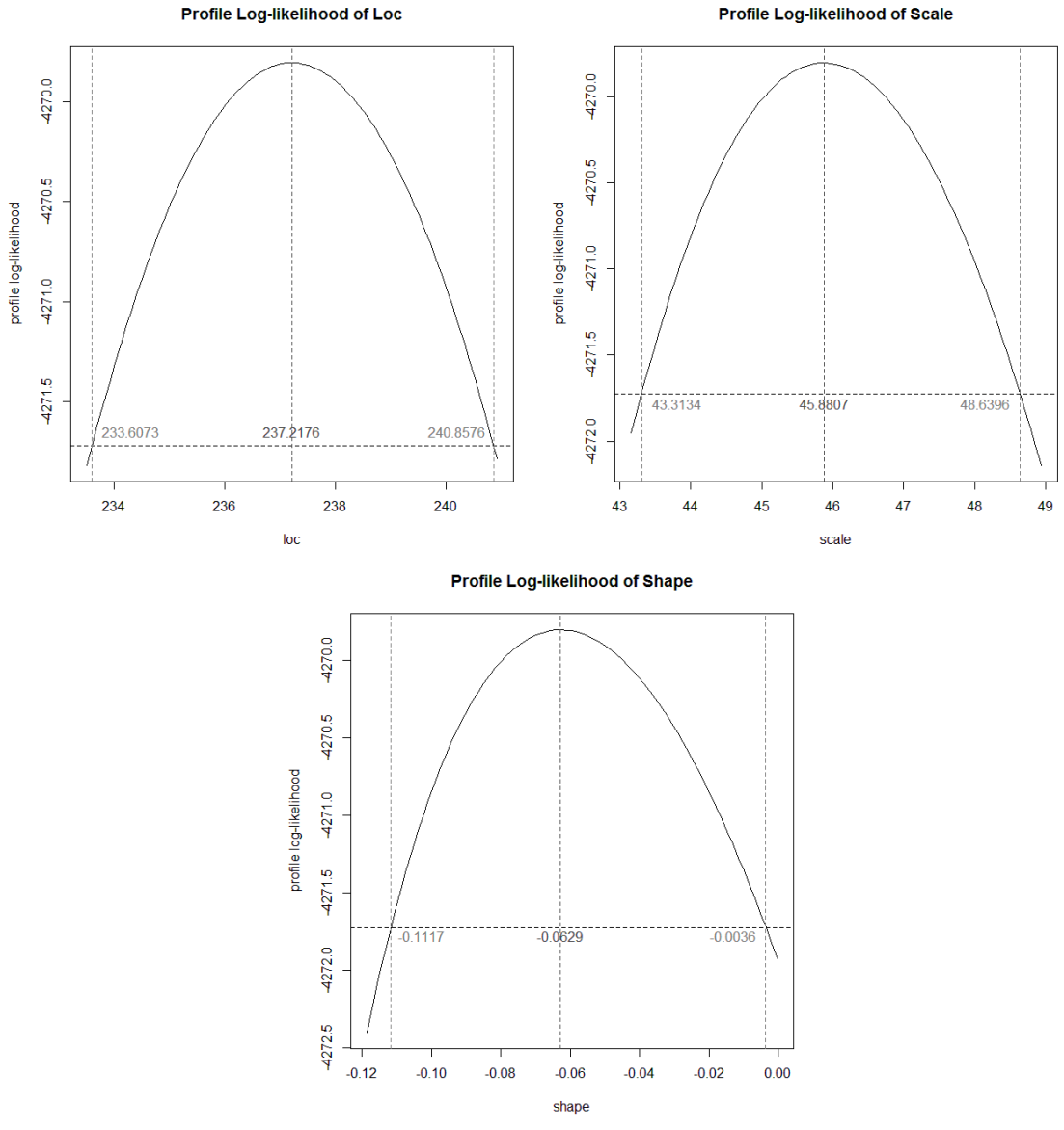

Figure 4.18: Profile Log-Likelihood plots and 95% CI's for location (top-left), scale (top-right) and shape (bottom) parameters of the GEV fit to the female SA freedivers' individual best records

After this analysis and now regarding the GEV fit to the maxima sample of the personal

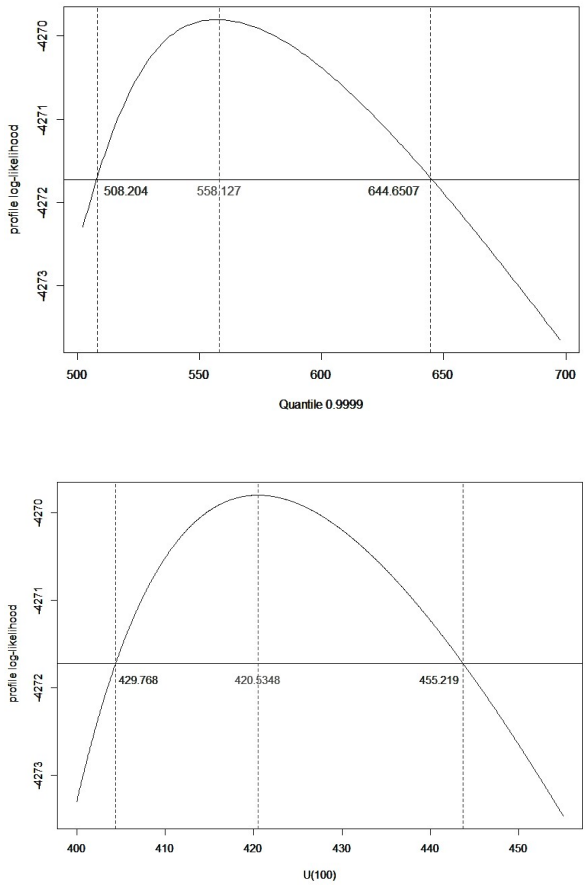

Figure 4.19: Profile Log-Likelihood plot and 95% CI for $\chi_{0.0001}$ (top) and $U(100)$ (bottom) under the GEV fit to the female SA freedivers' individual best records

records of female SA freedivers, under the current stationarity setup, we can estimate:

- The probability that a female freediver will set her best personal SA record above the current world record of 9 minutes and 2 seconds (542 seconds) is at best approximately 0.018%, which is even more reduced than estimated previously;
- There is approximately a 0.01% probability that that a female freediver will set her best personal SA record above $\approx$ 9 minutes and 18 seconds (558 seconds), meaning this is a very unlikely mark;
- In average, about 100 female SA freedivers must set their best mark ever so that a best individual record above $\approx$ 7 minutes (420 seconds) is observed;
- The maximum apnea time a female SA freediver can possibly set as her personal record is $\approx$ 16 minutes and 8 seconds (968 seconds).

We also came to the conclusion that a distribution function from the Weibull-domain is more suitable for modeling our data than a Gumbel-type distribution

4.1.1.2 Peaks Over Threshold Method

For the application of the Classical Extremes Method – the Gumbel approach – we looked at our observations as each representing the maximum of a single block we could not access. Now, in order to apply the Peaks Over Threshold methodology and to guarantee the results are comparable to the ones presented in the previous section, we must look at our sample of size $n = 795$ personal best records as a realization of the i.i.d. sample $(X_1, \ldots, X_n)$. As such, our variable X with d.f. F represents the best personal record of a female SA competitive freediver (with the definition of a competition mark as set in the introduction of the current Chapter – records over 3 minutes). Again it will not be considered here any temporal evolution or influence on the data, the resulting stationary inference regarding simply the current state-of-the-art.

We will rely on the methodology presented in section 3.1.2 for trying to fit a Generalized Pareto distribution to the N_u-sized sample of excesses above a threshold u to be determined, estimating the corresponding shape and scale parameters (ξ, σ_u). Having set the desired threshold, we will work with the observed excesses sample $(y_1, \ldots, y_{N_u})$ computed as $Y|Y > 0$ with $Y := X - u$, where N_u is the number of best personal records from the original sample that exceed the apnea time u.

Choice of Threshold and Preliminary Analysis Before applying the procedures of the POT approach there are two tasks we should deal with. The first very relevant point is finding the appropriate threshold u above which we will select the observations for the methodology. The importance of this choice has been stressed many times in the previous Chapters. Once the suitably hight level to be set has been defined, then it will be necessary a first assessment of the tail behaviour of our distribution, similarly to what has been done for the Block Maxima approach.

In section 3.1.2.6 we presented a pragmatic methodology proposed by Davison and Smith (1990) for choosing an appropriate threshold for the POT method, based on the plot of the sample mean excess plot. Recall this technique consists in finding the point on the ME-plot such that a linear pattern is visible to its left. Since we are dealing with already relatively high values of apnea time, above 3 minutes (180 seconds), we would expect a suitable threshold to be located around the 4 or 5 minute marks. Moreover, as stated before, a 4 minute breath hold is already physiologically a considerable achievement.

In Appendix A.11 can be found the **R** code written for the plotting of the empirical ME-plot regarding the sample at hand. For a better visualization of the patterns in the sample path, and with the objective of obtaining better fitted straight lines posteriorly, the last three points of the empirical m.e.f. were omitted from the plot, since they were disruptive to its analysis.

The ME-plot was reproduced twice, as seen in Figure 4.20, with the intention of comparing the linear fits to the left of both considered points: 5 minutes (300 seconds), the most obvious point in the plot where it appears to happen a change in trend for the function, and 4 minutes (240 seconds), where a smaller inflection happens, but to the left of which the linear pattern is very evident.

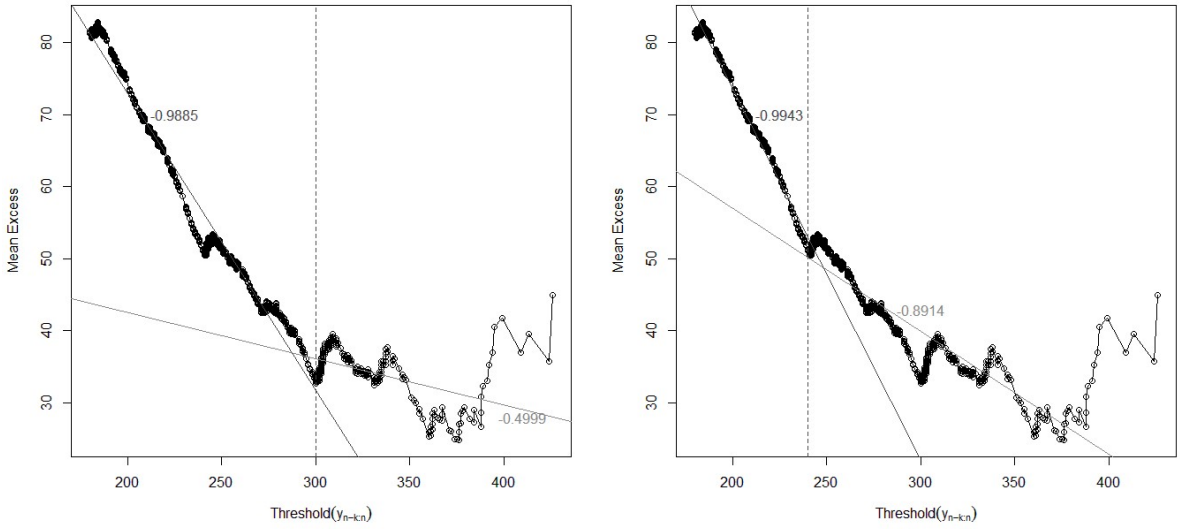

Figure 4.20: Sample ME-plot of the female SA freedivers' individual best records and corresponding linear fitting for thresholds of 5 minutes (left) and 4 minutes (right)

When considering the 300 seconds threshold, in the left plot of Figure 4.20, we can see that to the linear regression line fitted to the left of this point corresponds a correlation of absolute value 0.9885. This is very satisfying for a linear fit, but the path to the right of this point is too irregular, leading us to consider the possibility of this being too high of a threshold. The correlation corresponding to the linear fit to the right of the 300 seconds threshold has a low absolute value of approx. 0.5. Looking now at the plot on the right of the same Figure, we see that the fitted line to the left of the threshold point of 240 seconds corresponds to a correlation of -0.9943, even more close in absolute value to 1 than for the 300 seconds threshold. As such, we have that the sample path to the left of 240 is more approximately straight than when considering all the points up to 300, suggesting that 240 is a sufficiently high threshold. The linear fit to the right of this value also presents a higher absolute correlation than the corresponding fit above 300, of approximately 90%.

Having now both the physiological and ME-plot arguments in favor of setting the 240 seconds mark as the threshold for the method, we seek a final confirmation from the plots in Figure 4.21, which were produced by the last line in the code from Appendix A.11, and which represent the Maximum Likelihood estimates for the parameters of a GP distribution fitted to excesses over each considered threshold. Since we aim for the best and most accurate estimates possible, a suitable threshold will be located in a stable region for the estimates, indicating that small changes in the level considered should not greatly affect the results of the POT inference.

We can see a fairly stable area in both plots for values of the threshold roughly between 200 and 280. The confidence intervals for the parameters, also represented, are in this region noticeably tight, meaning very precise estimates. As such, a threshold level around 250 would here appear to be a good choice, specially when compared with our other alternative choice of 300 seconds.

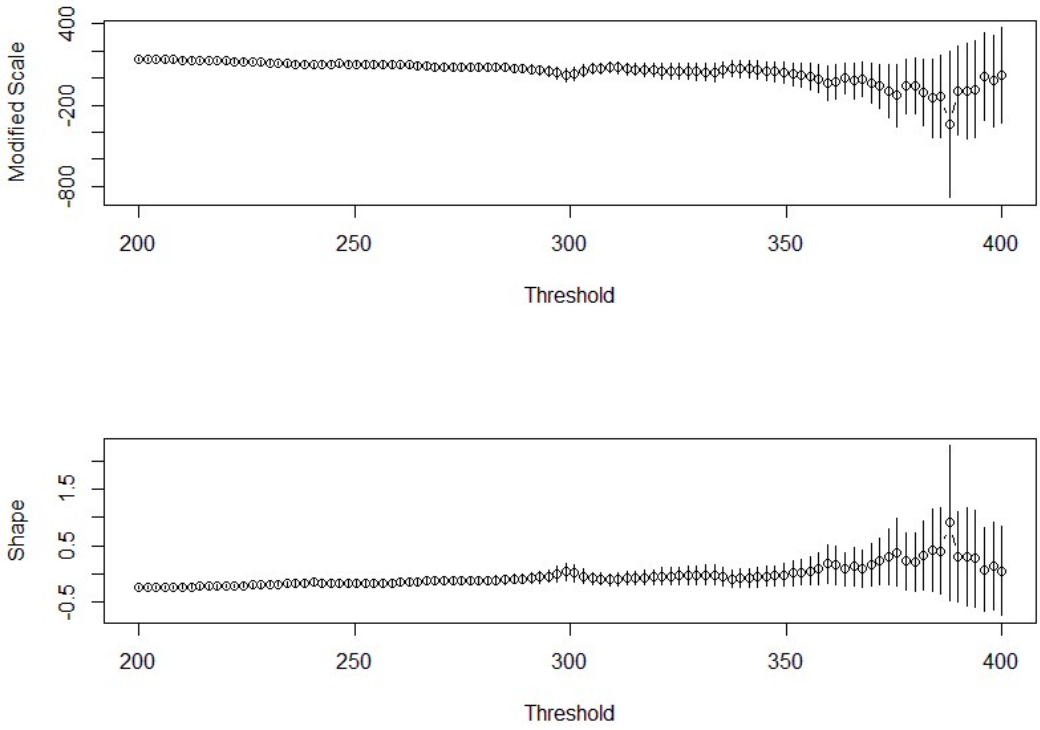

Figure 4.21: Parameter estimates for a GP fit to the to the excesses over a threshold of the female SA freedivers' individual best records, for each threshold

Around the latter, there is a slightly pronounced change in the estimates for the parameters, as well as a growth in the size of the confidence intervals. This information, associated with the previously referred arguments, makes us comfortable in declaring the 4 minutes mark as the threshold to be used in our application of the POT methodology.

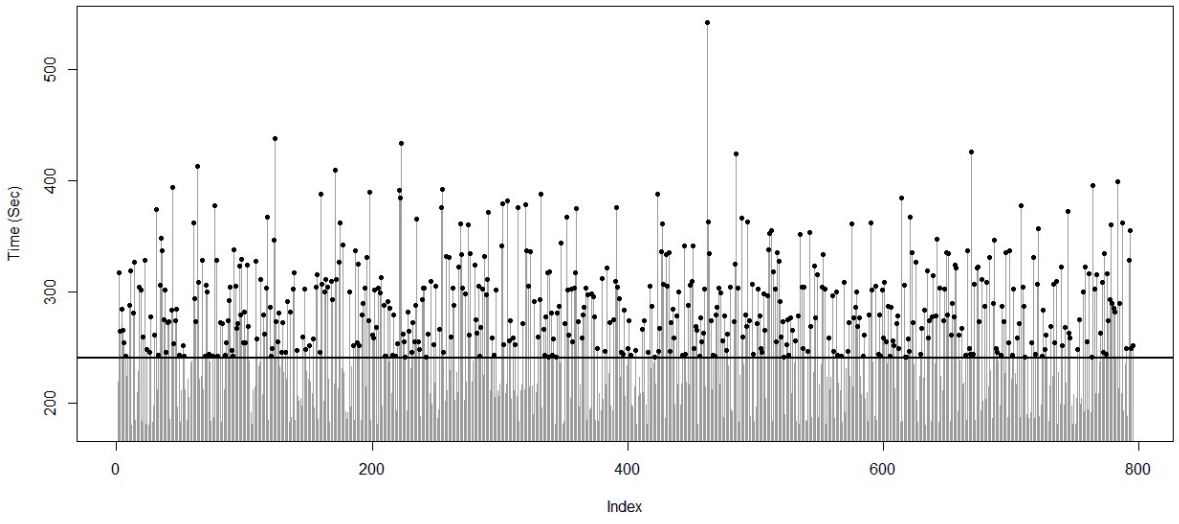

Figure 4.22: Exceedances over $u = 240$ seconds of the female SA freedivers' individual best records

Taking this threshold will reduce the dimension of our sample. But since we had a very large sample of 795 elements, and are left with 515 observations that fall above $u = 240$ (a still considerably large sample size), this dimensionality reduction does not pose a problem to the validity of the POT approach to this case. In Figure 4.22 we see represented all our 795 initial observations, the threshold line of 240 seconds and emphasised by the black dots all of the 515 exceedance observations (corresponding code in Appendix A.12).

Now that the threshold has been set, we can specify the sample of excesses at hand as $(y_1, \ldots, y_{515})$ from the r.v. $Y|Y > 0$ with $Y := X - 240$. This is the sample to which we will try to fit a GP distribution. From the application of the Gumbel method in the last section (and having in mind the Pickands-Balkema-de Haan Theorem – Theorem 2.4.1 in this dissertation) we learned that it is most likely that we will obtain a negative estimate for the EVI, meaning that the right tail of our data's underlaying distribution function F should be lighter than an exponential tail, bounded. That is if the assumptions made about our data are indeed correct, being that we still have not checked the validity of the stationarity assumption.

Let us then procure some insight on the sample's behaviour through the same preliminary tool used before: the quantile-quantile plots. The $\xi = 0$ case corresponds, as we know, to the frontier Gumbel domain of attraction, and the excesses over suitable thresholds obtained from variables whose distributions belong this domain are expected to be of the Exponential type – the particular case of the GP distribution for null EVI.

Hence, we first analyze the Exponential qq-plot of our excesses data, specifically the possible linear relation, or lack there off, between the theoretical Exponential quantiles and the order statistics $y_{i:N_u}$. A technical detail arises from the fact that the straight line relating the theoretical Exponential quantiles intercepts the ordinate axis at 0, and as such the regression straight line relating the theoretical with the empirical quantiles must be fitted with a null intercept as well. The code that produces the qq-plot in Figure 4.23 is comprised in Appendix A.13.

We can see in the right half of the plot the sample path of the excesses deviating considerably from the linear pattern. Although the correlation between the empirical and theoretical quantiles is around 98%, we suspect that our sample of excesses will not be properly fitted by an Exponential distribution. Also, the apparent convexity in the plot points, as expected, to a lighter than exponential tail, i.e., to a better fitted GP distribution with negative tail index.

A GP qq-plot is obtained by plotting the empirical quantiles $y_{i:N_u}$ (sorted excesses) against the theoretical GP quantiles $\frac{(1-p_i)^{-\xi}-1}{\xi}$ where p_i are again the plotting positions we define as $p_i := {}^{i}/_{(N_u+1)}$. We will go around the problem of the unknown EVI in a similar fashion as before, resorting to the correlation plot to get a preliminary estimate for ξ, as suggested by Beirlant et al. (2004). This plot is constructed with the successive correlations computed between the empirical quantiles $y_{i:N_u}$ and the theoretical quantiles of a GP distribution with shape parameter varying from -1 to 0.5. The code in Appendix A.14 yields the correlation plot in Figure 4.24, where the maximum correlation and corresponding value of the EVI were marked.

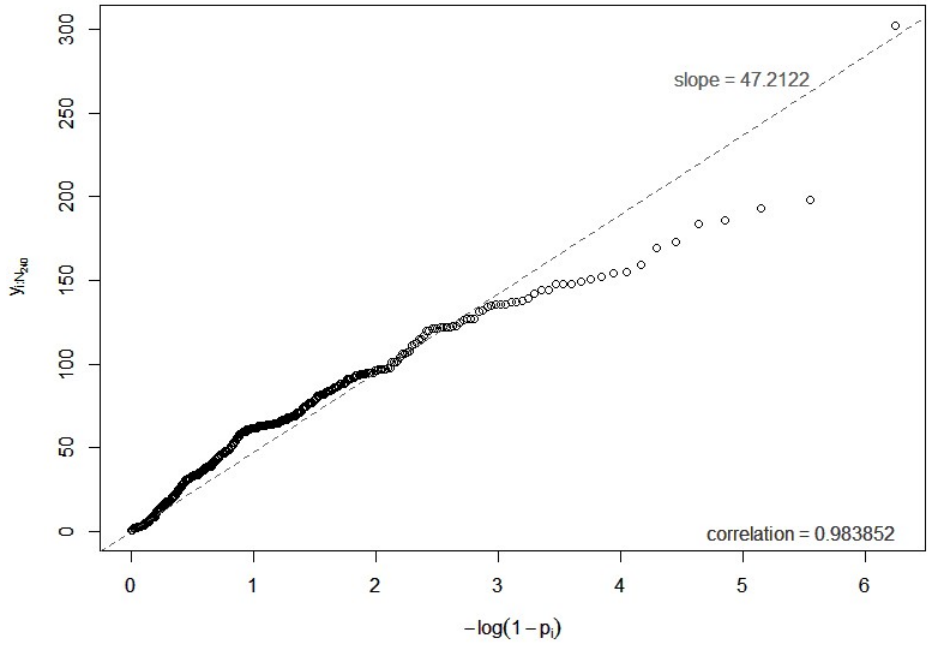

Figure 4.23: Exponential QQ-plot of the excesses over $u = 240$ seconds of the female SA freedivers' individual best records

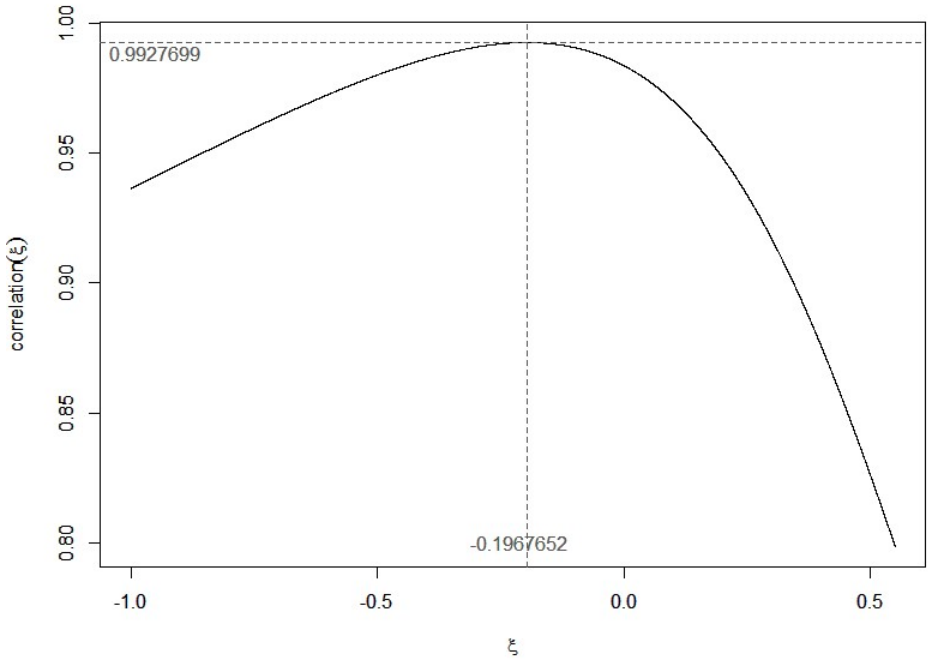

Figure 4.24: Correlation plot for the GP family QQ-pot of the excesses over $u = 240$ seconds of the female SA freedivers' individual best records

Considering then the optimal value of the EVI as $\hat{\xi} = -0.1967652$ in the sense that it yields the largest (high) correlation of approximately 99.3% for the fit, we have the corresponding qq-plot, which shows a much more extended linear pattern, indicating a better fit of the GP distribution with that EVI to the excesses over 240 seconds sample (corresponding code in Appendix A.15). There is only one significant deviation from linearity corresponding to the sample maximum, as previously stated, due to the very large magnitude of that value. Once again, the Least Squares line had to be fitted without intercept.

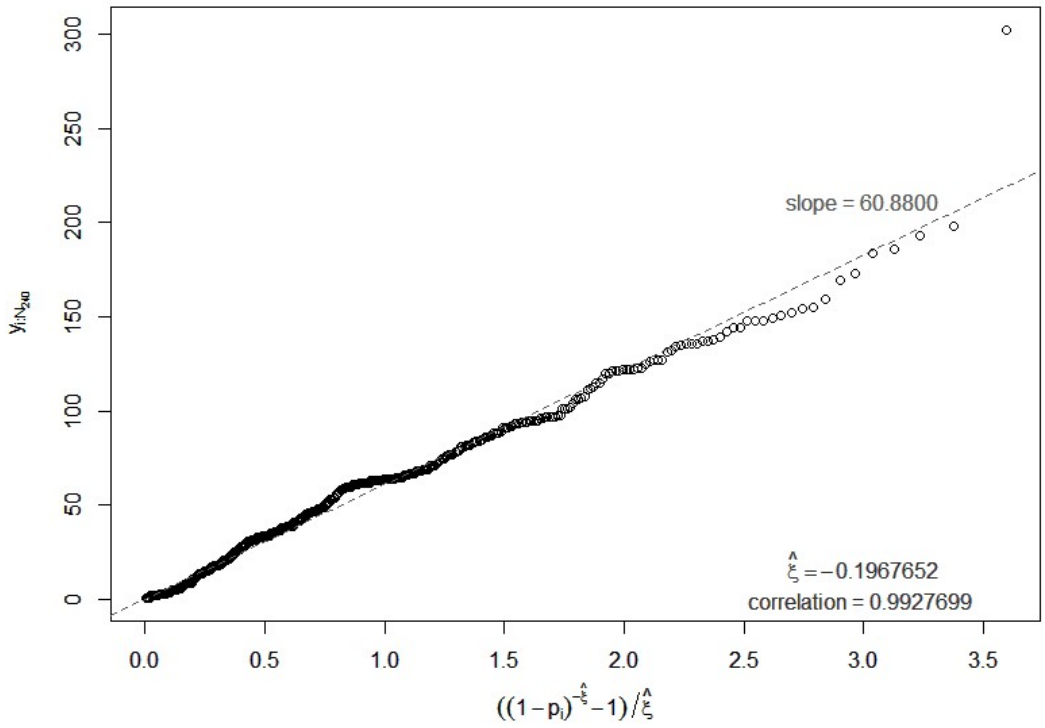

Figure 4.25: GP QQ-plot for $\hat{\xi} = -0.1967652$ of the excesses over $u = 240$ seconds of the female SA freedivers' individual best records

This preliminary fit presents us with the estimate $\hat{\sigma_u} = 60.88$ for the scale parameter of the GP distributions with comfortably negative estimated EVI $\hat{\xi} = -0.1967652$. Unlike what happened with the preliminary analysis performed for the BM approach, very few doubts remain in this case regarding what type of distribution can better adjust to the $(y_1, \ldots, y_{515})$ sample. The Exponential fit was not nearly as satisfactory as the GP fit, and the estimated tail index is not very close to 0, making us lean towards concluding that F belongs to the Weibull max-domain. However, in the next section we will perform some objective statistical tests for choosing the domain of attraction which, based on each Generalized Pareto model suitability for the excesses, F is most likely to belong to.

Statistical Choice of Domain of Attraction Recalling now the testing procedures presented in section 3.1.2.5, we will try to consolidate and strengthen our preliminary choice of the GP distribution with negative EVI for modelling the excesses above the chosen threshold $u = 240$. The objective is to fully discard the Exponential distribution as a suitable candidate for Y, corresponding to $\xi = 0$, in favor of a GP distribution with $\xi \neq 0$ (through a two-sided test) or even preferably $\xi < 0$ (through a test with a one-sided alternative). Thus we will apply the testing procedures for the following sets of hypothesis:

$$H_0 : \xi = 0 \qquad versus \qquad H_1^{(1)} : \xi \neq 0 \tag{4.3}$$

and

$$H_0 : \xi = 0 \qquad versus \qquad H_1^{(2)} : \xi < 0 \, . \tag{4.4}$$

The test statistics on which these procedures are based are constructed not on the sample of excesses $(Y_1, \ldots, Y_{N_u})$ but on the sample of the exceedances above u, that is, the real value of the original observations larger than 240 seconds. We denote, as before, this i.i.d. sample as $(W_1, \ldots, W_{N_u})$ and test the Exponential fit to it. Discarding the suitability of an Exponential distribution to the r.v. of the exceedances W is equivalent to doing so for the variable of excesses Y, since they are related by the simple reparameterization $w = y + u$.

All the code referring to this section can be found in Appendix A.16. Since one of the tests to be applied is the Likelihood Ratio test, it was required the prior Maximum Likelihood estimation of the GP distribution's parameters, for now performed with resource to the functions in the **R** package *fitdistrplus*, first for an Exponential fit, and then considering a distribution from the unrestricted GP family. The obtained estimates are compiled in Table 4.8.

Table 4.8: ML estimates from the *fitdistrplus* package for the scale and shape parameters for the fitted Exponential and GP distributions to the excesses over $u = 240$ seconds of the female SA freedivers' individual best records.

	$\hat{\sigma_u}$	$\hat{\xi}$
Exponential	51.39848	-
GP	59.28822	-0.16097

Firstly, let us discuss Table 4.9 where we find the test results for the hypothesis with the two-sided alternative (4.3), at the usual significance levels of 5% or 1%.

The test results are very clear, unlike what happened when testing for the Gumbel fit in the previous approach. Both p-values for the LRT and for the test suggested by Marohn (2000) are very close to 0, significantly smaller than the smallest usual significance level of 0.01, and as such lead in both procedures to the clear rejection of null hypothesis corresponding to an Exponential-type tail of the underlying distribution to our data. However, since the alternative hypothesis is two-sided, these tests don't allow us to conclude if a negative or positive EVI would be appropriate. Of course we already suspect a negative EVI, and have been consistently dismissing

Table 4.9: Test results for hypothesis in (4.3) for the excesses over $u = 240$ seconds of the female SA freedivers' individual best records.

Test	Observed Statistic	Observed p-value	Decision
LR	18.41124	1.78005e-05	Rejection of H_0 at every usual significance level
T_{N_u} **Statistic**$^{(*)}$	-3.702529	0.0002135	Rejection of H_0 at every usual significance level

(*) as in Marohn (2000)

heavy tailed distributions as suitable adjustment candidates, so the following procedures aim to test the one sided hypothesis in (4.4), corresponding to a light-tailed GP fit to the excesses. The test results figure in Table 4.10.

Table 4.10: Test results for hypothesis in (4.4) for the excesses over $u = 240$ seconds of the female SA freedivers' individual best records.

Test	Observed Statistic	Observed p-value	Decision
T_{N_u} **Statistic**$^{(*)}$	-3.702529	0.0001067	Rejection of H_0 at every usual significance level
G_{N_u} **Statistic**$^{(**)}$	-4.916655	5.002582e-60	Rejection of H_0 at virtually any significance level

(*) as in Marohn (2000)

(**) as in Gomes and van Monfort (1986)

The decisions suggested by these tests are coherent with the analysis made so far, in the sense that they reject the Exponential fit in favor of a lighter tailed Generalized Pareto distribution for every usual significance level applicable. In fact, the p-value associated with the test based on the Gomes and van Monfort (1986) statistic is statistically 0, meaning the rejection of H_0 is absolute.

In section 3.1.2.5 another type of tests were suggested – the goodness-of-fit tests, similar to the ones already performed in the previous section, that are based in the proximity of the empirical and theoretical distribution functions. We will use the Kolmogorov-Smirnov test as specified by Lilliefors (1969) to test the goodness-of-fit of the Exponential distribution – Table 4.11 –, and the Cramér-von Mises and Anderson-Darling statistics for testing the goodness-of-fit of a GP distribution – Table 4.12. For both cases were used the ML estimates of the models parameters. As before, the corresponding code can be found in Appendix A.16.

Note that the critical value Tables 3.4 and 3.5 regarding the Cramér-von Mises and Anderson-Darling statistics should be consulted at the estimated value of the EVI, when unknown, but our estimate is $\hat{\xi} = -0.16097$ not contained in said tables.

Table 4.11: Goodness-of-fit of the Exponential distribution test results for the excesses over $u = 240$ seconds of the female SA freedivers' individual best records.

Test	Observed Statistic	Critical Point 1%	Decision
Kolmogorov-Smirnov	0.09562227	0.05508158	Rejection of the Exponential model

Table 4.12: Goodness-of-fit of the GP distribution test results for the excesses over $u = 240$ seconds of the female SA freedivers' individual best records.

Test	Observed Statistic	Critical Point 1% for $\xi = -0.1/-0.2$	Decision
Cramér-von Mises	0.4338382	0.210/0.200	Rejection of the GP model
Anderson-Darling	2.536057	1.348/1.296	Rejection of the GP model

To solve this problem it is suggested in the referred paper from Choulakian and Stephens (2001) that the tables be entered at the closest EVI value possible. We then considered both critical values for the EVI set as -0.1 and -0.2.

The decision suggested by the Kolmogorov-Smirnov test is concordant with the other tests previously performed in this section: we should reject at significance levels no lower than 1% that the Exponential distribution would be a good fit for our excesses data. However, both goodness-of-fit tests for the GP distribution with ML estimated parameters conclude on the rejection of said model (for both EVI values at which we entered the critical values tables, thusly the conclusion being valid for our intermediate $\hat{\xi}$). This conclusion contradicts the analysis performed so far, and we suspect this problem derives from our sample being too large, increasing the power and sensitivity of the tests, causing that even small deviations of the empirical from the theoretical d.f. could lead to the rejection of H_0. Also the alternative hypothesis in these goodness-of-fit tests comprises any model different from the one set in H_0. In the previous ones the scenario is more restrict, as we are faced with a dilemma between Exponential versus Beta model, in the GP family.

We are comfortable, after this analysis, in affirming that a GP distribution with a negative EVI is the best fit for our sample of excesses over 240 seconds. Despite this conclusion, in the following section we will still analyse in detail the Exponential fit and perform inference under this assumption, not for thinking it is appropriate, but only for keeping in mind that the Block Maxima methodology was incapable of fully discarding the attraction of our extreme data the Gumbel max-domain.

Estimation of Parameters and Other Extreme Value Indicators Analogously to what was done under the Gumbel and GEV fits for the maxima sample, in the last section, we will now

fit an Exponential ($\xi = 0$) and a GP ($\xi \neq 0$) distributions to the sample of the excesses above $u = 240$ seconds of the female SA freedivers' individual best records, aiming at the inference on the same indicators as before: $P[X > 542]$ – the probability that a female freediver will set her personal best record above the current world record; $\chi_{0.0001}$ – the individual female best competitive Static Apnea time that is registered with a 0.01% probability; $U(100)$ – the apnea individual record time that is exceeded, on average, once every 100 best personal records set; when possible, the finite x^F – absolute maximum apnea personal record time that could possibly be set by a SA competitive female freediver, only computed when the previous estimation of the EVI produces a negative value. The functional expressions for this parameters, under the POT methodology and dependent on the estimation of (ξ, σ_u) the models' core parameters, can be seen in section 3.1.2.3.

For the Exponential fit, two estimation methods were used – the preliminary estimation based on the qq-plot in Figure 4.23 and the Maximum Likelihood estimation approached in section 3.1.2.1. Regarding the GP fit, to these two methodologies (the preliminary one being now based on the qq-plot in Figure 4.25) was added the Probability Weighted Moments estimation presented in section 3.1.2.2. For this case, and again with the objective of verifying that the specific programming of each function did not interfere significantly with the estimation results, for the ML method, we made use of pre-programmed fitting functions from 4 different packages in the **R** software: *fitdistrplus* (already employed in the previous section on the statistical testing), *evir*, *ismev* and *evd*. Having found the corresponding estimates for the models' shape and scale parameters, those were used in computing estimates for the other indicators of interest mentioned.

Table 4.13 shows all the estimates obtained for the Exponential fit to $(y_1, \ldots, y_{515})$ through the preliminary and ML estimation methods, corresponding to the output of the code in Appendix A.17.

Table 4.13: Estimates for the Exponential fit to the excesses over $u = 240$ seconds of the female SA freedivers' individual best records.

Method	Package	$\hat{\sigma}$	$\widehat{P[X > 542]}$	$\widehat{\chi_{0.0001}}$	$\widehat{U(100)}$
Preliminary	-	47.2122	0.001079965	654.3421	436.9218
ML	*fitdistrplus*	51.39848	0.001818333	691.0816	454.3828

The first evident aspect is the significant difference between the preliminary and ML estimates for all the estimated parameters. This could mean the ML method improves slightly the poor Exponential fit observed in the preliminary analysis. However, we know from the statistical tests for the model choice that said improvement is still not enough to consider the Exponential distribution as well adjusted to our excesses data. The preliminary estimates $\widehat{P[X > 542]}$, $\widehat{\chi_{0.0001}}$ and $\widehat{U(100)}$ are in fact very close to the respective ML estimates under the Gumbel fit to the maxima sample. On the other hand, the ML estimates under this Exponential fit for the same parameters are the highest obtained so far, considering all the BM estimation. But, as stated before, we have serious doubts that an Exponential fit would be appropriate, so we do not greatly mind these estimates.

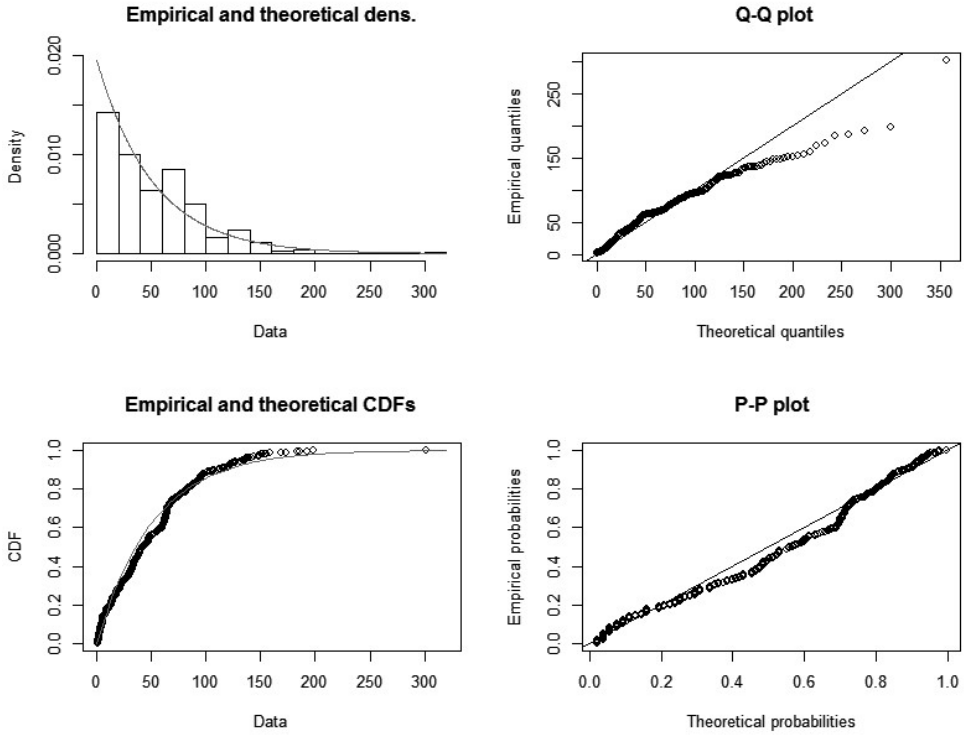

Figure 4.26: Diagnostic plots given by the *fitdistrplus* package for the Exponential fit to the excesses over $u = 240$ seconds of the female SA freedivers' individual best records

Our doubts are increased by the diagnostic plots in Figure 4.26, from the *fitdistrplus* package for the Exponential fit to the excesses. Besides the very clear convexity of the qq-plot sample path (top-right plot), we can also see a bigger dissonance between the empirical and theoretical d.f. than the observed so far (bottom-left plot) and also a poor linearity in the probability-probability plot.

Let us briefly mention the 95% Confidence Intervals based on the profile log-likelihood function that were computed for the scale parameter of the fitted Exponential distribution, as well as for the indicators $\chi_{0.0001}$ and $U(100)$. These correspond to the plots in Figures 4.27 and 4.28, and are estimated as

$$CI^{95\%}_{\sigma_u}(H_0) = [47.20310\,;\, 56.10217]\,,$$

$$CI^{95\%}_{\chi_{0.0001}}(H_0) = [654.2622\,;\, 732.3882]\,,$$

$$CI^{95\%}_{U(100)}(H_0) = [436.8839\,;\, 474.0144]\,.$$

As we have mentioned, not much attention will be paid to this estimation, because of the poor fitting of the Exponential distribution, and as such we will at once skip to the next phase: the fitting of the Generalized Pareto distribution to the sample of excesses over $u = 240$ seconds of the female SA freedivers' individual best records.

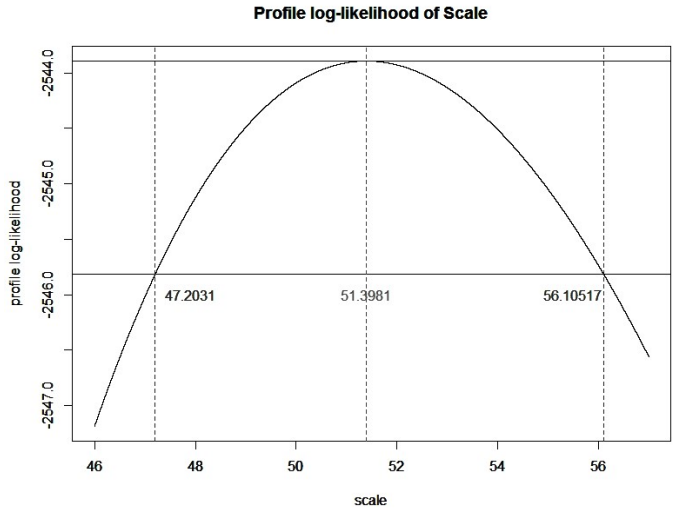

Figure 4.27: Profile Log-Likelihood plot and 95% CI's for the scale parameter of the Exponential fit to the excesses over $u = 240$ seconds of the female SA freedivers' individual best records

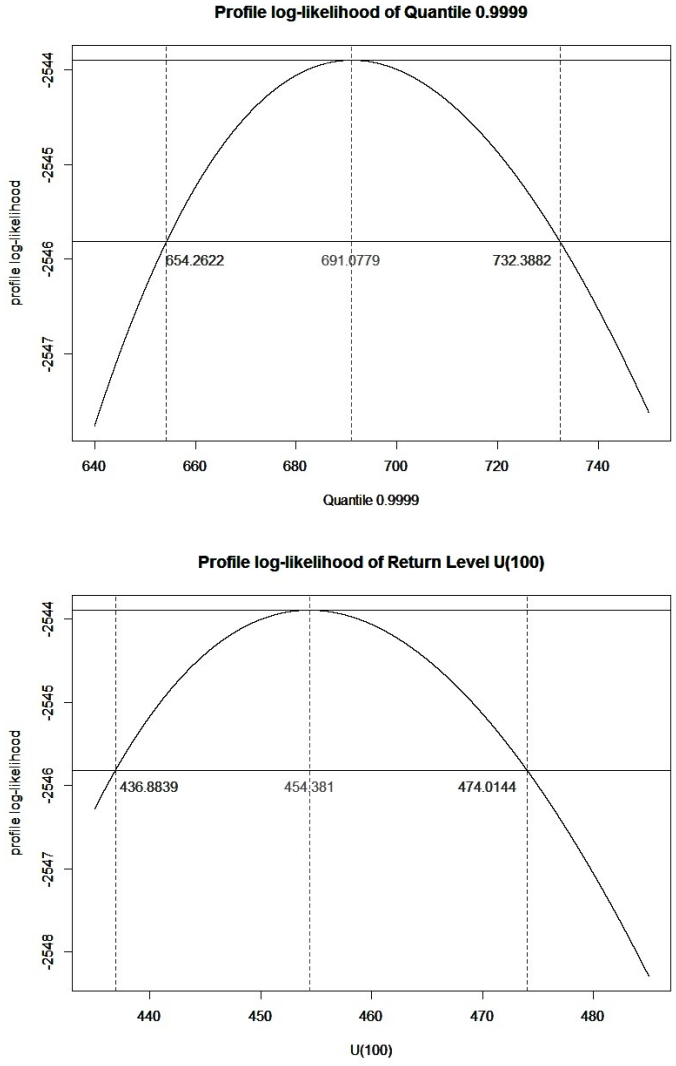

Figure 4.28: Profile Log-Likelihood plot and 95% CI for $\chi_{0.0001}$ (top) and $U(100)$ (bottom) under the Exponential fit to the excesses over $u = 240$ seconds of the female SA freedivers' individual best records

We are guided again by the methodology in section 3.1.2, but now focussing on the presented expressions regarding the case $\xi \neq 0$. Now resorting to preliminary, ML and also PWM estimations, we comprised in Table 4.14 the obtained estimates for the GP model's shape and scale parameters $(\hat{\xi}, \hat{\sigma_u})$ and for the mentioned indicators $P[\widehat{X > 542}]$, $\widehat{\chi_{0.0001}}$, $\widehat{U(100)}$ and, for the cases when $\hat{\xi} < 0$, also $\widehat{x^F}$. The code developed in the **R** software for this GP fitting and all the subsequent analysis and plots can be found in Appendix A.18.

Table 4.14: Estimates for the GP fit to the excesses over $u = 240$ seconds of the female SA freedivers' individual best records.

Method	Package	$\hat{\xi}$	$\hat{\sigma}$	$P[\widehat{X > 542}]$	$\widehat{\chi_{0.0001}}$	$\widehat{U(100)}$	$\widehat{x^F}$
Prelim.	-	-0.1967652	60.8800	3.740943e-09	494.3784	413.2291	549.4043
ML	*fitdistrplus*	-0.1609684	59.2882224	1.534186e-05	518.6399	420.1119	608.3221
	evir	-0.1615137	59.2839675	1.441435e-05	518.1059	419.9170	607.0521
	ismev	-0.1607758	59.2776040	1.558913e-05	518.7718	420.1439	608.6972
	evd	-0.16088	59.29087	1.551055e-05	518.7358	420.1495	608.5411
PWM	*evir*	-0.2506255	64.2797231	**NaN**	468.0454	406.3085	**496.4772**

Evident from a first look at Table 4.14 is that all three estimation methods provide negative estimates for the tail index ξ, as we expected. Completely out of question is then an underlying distribution F from the Fréchet domain of attraction, that is, a heavy-tailed distribution.

Analysing the preliminary estimates, obtained from the GP qq-plot in Figure 4.25, we see the EVI estimate is close to -0.2, significantly negative and implying a right endpoint estimate of approx. 549 seconds, only about 7 seconds higher than the sample maximum. So these estimates are unlikely to be accurate.

Moving to the PWM estimates, we see some problems with the estimation. Firstly, the EVI estimate is the most negative obtained so far, around -0.25, which translates in a very short tail and produces another problem – the right endpoint comes estimated below the value of the sample maximum, which is obviously absurd. Moreover, the computation of the exceedance probability of the sample maximum produces, consequently, the *Not a Number* error. In truth, if the right endpoint is lower than 542, the exceedance probability of this value should be estimated as 0. Because of this problems, we will discard the complete PWM estimation for this GP fit.

Thus, we are left with the estimates given my the ML method, which are once again extremely close for all the packages used, proving the choice of function between the ones suggested for the ML fitting is not determinant. This methodology also produces, as stated before, negative estimates for the EVI, but not so negative as the ones from the other methods, being around -0.16. However, this estimate is still much smaller than the ones obtained for the EVI under the GEV fit for the maxima sample, which explains the also lower estimate for the endpoint of around 608 seconds, and the very low exceedance probability estimated for the sample maximum around 0.00001 (about 0.001%). An interesting fact is that although the estimated extremal quantile $\widehat{\chi_{0.0001}}$ comes almost 40 seconds lower than when estimated under the GEV fit to the maxima sample, both estimates for $U(100)$ are almost identical. This shows the growing influence

the fitted models have towards the right end of the sample/distribution, thus illustrating the importance of EVT to the adjustment of really extreme values, not minding so much the more central data.

To comment on the quality of this adjustment we can once more analyse the graphical diagnostic tools from the *fitdistrplus* and *evd* packages, for they present complementing information, represented in Figures 4.29 and 4.30, respectively. Note that this evaluation regards the fit of the distribution with its parameters estimated by the Maximum Likelihood method. Again some interest lays in comparing the quality of this GP adjustment to the already assessed poor quality of the Exponential adjustment, as shown by Figure 4.26.

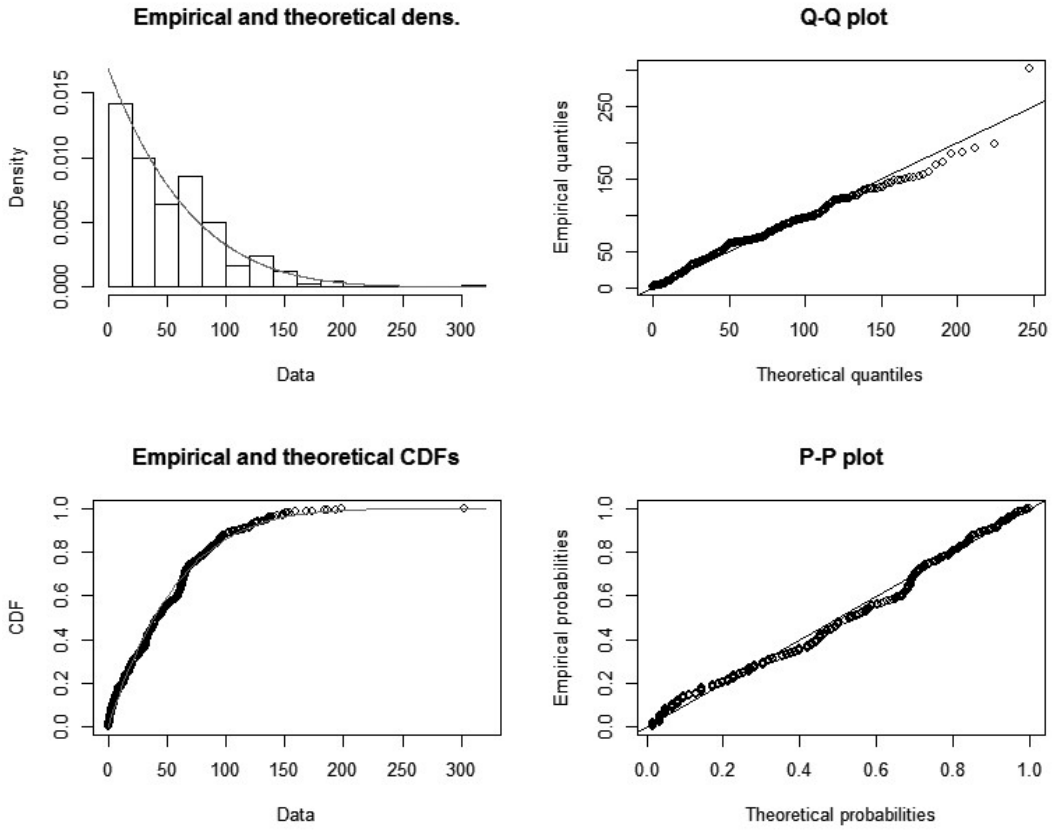

Figure 4.29: Diagnostic plots given by the *fitdistrplus* package for the GP fit to the excesses over $u = 240$ seconds of the female SA freedivers' individual best records

The main difference is observable in the qq-plots, but this had already been verified in the early preliminary analysis. The plot comparing the empirical and theoretical d.f.'s for the GP fit (bottom-left of Figure 4.29) shows a much higher concordance than its counterpart for the Exponential fit (bottom-left of Figure 4.26), showing this is clearly a much better fit for our sample of excesses. We can also find a high level of concordance in the bottom-left plot of Figure 4.30, where the empirical and theoretical probability density functions are compared. Specially in right portion of the curves, that is, the right tails have very similar behaviour.

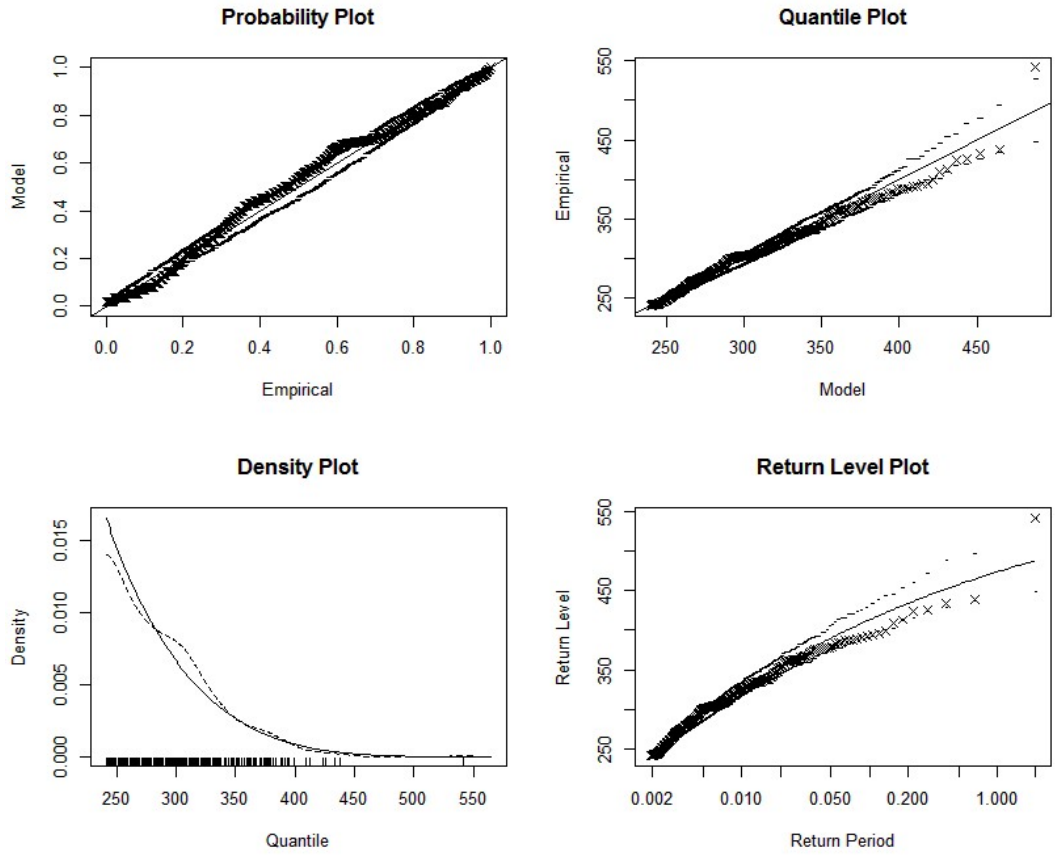

Figure 4.30: Diagnostic plots given by the *evd* package for the GP fit to the excesses over $u = 240$ seconds of the female SA freedivers' individual best records

Following once more the profile log-likelihood procedure already mentioned before, we computed 95% confidence intervals for the core parameters (ξ, σ_u) of the GP distribution. These were obtained with the help of the *evd* package from the **R** software (as exposed in Appendix A.18), as were the plots of the respective profile log-likelihood functions, seen here in Figure 4.31. Moreover, and as before, CI's for the indicators $\chi_{0.0001}$ and $U(100)$ were also constructed, and plotted the corresponding profile log-likelihood functions in Figure 4.32, but this time with the help of the *ismev* package. Considering then the ML estimates for the intrinsic parameters (ξ, σ_u), the 95% CI's based on the profile log-likelihood under the GEV fit to the data are

$$CI_{\xi}^{95\%}(H_{\hat{\xi}}) = [-0.1985433\,;\, -0.1024759]\,,$$

$$CI_{\sigma_u}^{95\%}(H_{\hat{\xi}}) = [53.7322532\,;\, 65.2997961]\,,$$

$$CI_{\chi_{0.0001}}^{95\%}(H_{\hat{\xi}}) = [496.75\,;\, 566.55]\,,$$

$$CI_{U(100)}^{95\%}(H_{\hat{\xi}}) = [408.59\,;\, 435.65]\,.$$

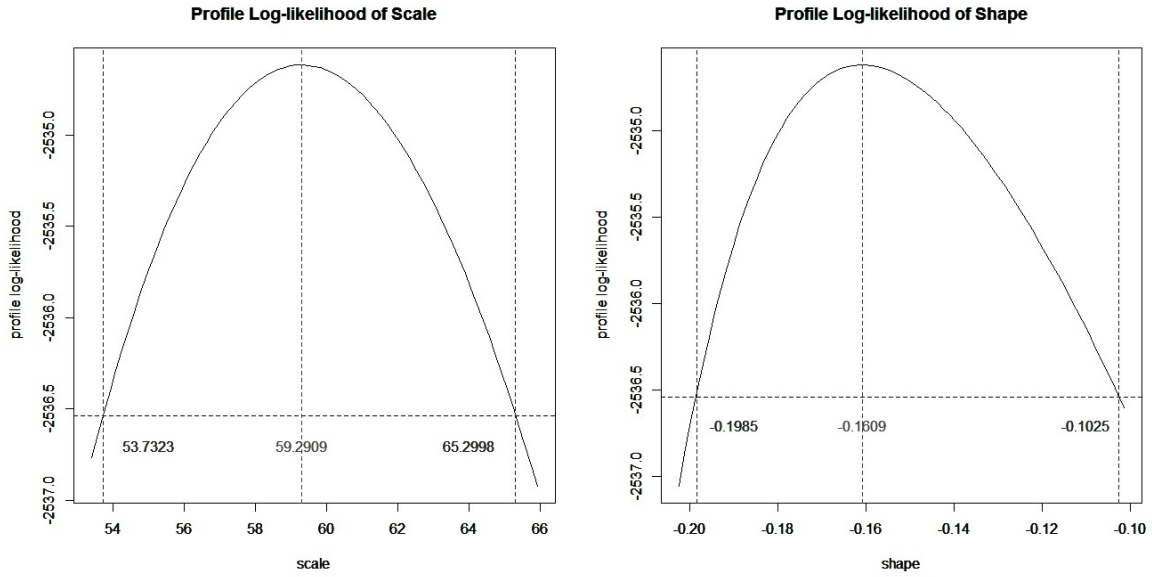

Figure 4.31: Profile Log-Likelihood plots and 95% CI's for scale (left) and shape (right) parameters of the GP fit to the excesses over $u = 240$ seconds of the female SA freedivers' individual best records

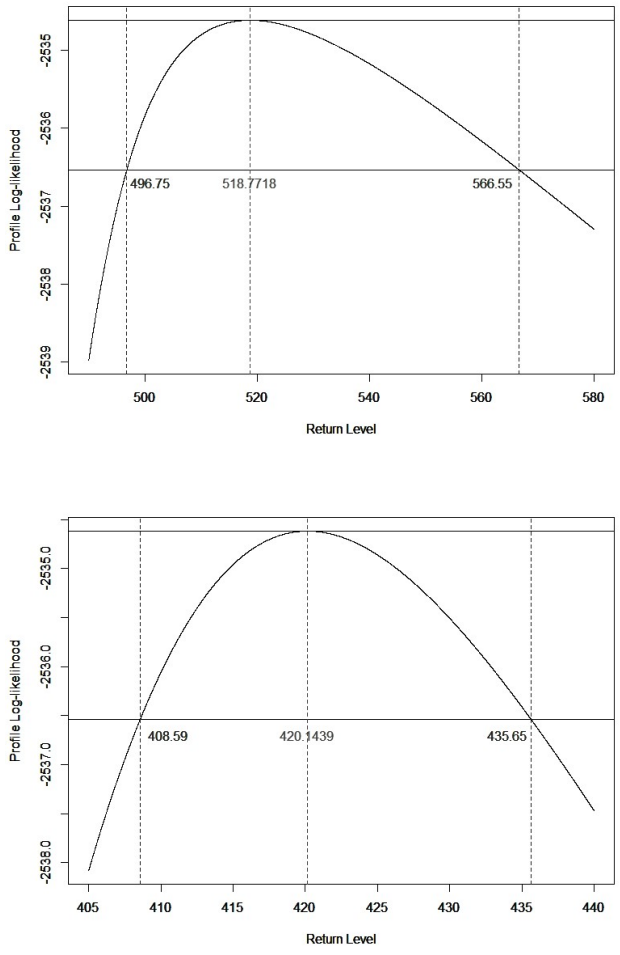

Figure 4.32: Profile Log-Likelihood plot and 95% CI for $\chi_{0.0001}$ (top) and $U(100)$ (bottom) under the GP fit to the excesses over $u = 240$ seconds of the female SA freedivers' individual best records

Note that the 95% CI for the EVI does not at all include the value 0, which means the Exponential distribution is not here included in the possible distributions for the adjustment with 95% confidence. Also note that the referred interval barely overlaps with the corresponding CI for the EVI under the GEV maxima estimation. The complete coverage of very negative values in this CI assures us of the lightness of the tail of the underlying distribution to our data, even if with the new information we now find it lighter than previously thought.

Regarding the $\chi_{0.0001}$ and $U(100)$ confidence intervals, it is once more clear that the profile log-likelihood based intervals are not necessarily centered around the ML estimation of the corresponding parameter. The CI's here computed have a smaller amplitude than their counterparts in the last section, this being more evident once again for $CI^{95\%}_{\chi_{0.0001}}(H_{\hat{\xi}})$ than for $CI^{95\%}_{U(100)}(H_{\hat{\xi}})$.

After this analysis of the GP fit to the sample of excesses over $u = 240$ seconds of the personal records of female SA freedivers, under the current stationarity setup, we can make the following estimation:

- The probability that a female freediver will set her best personal SA record above the current world record of 9 minutes and 2 seconds (542 seconds) is at best approximately 0.001%, which is even more reduced than previously estimated;
- There is approximately a 0.01% probability that that a female freediver will set her best personal SA record above $\approx$ 8 minutes and 38 seconds (518 seconds), meaning this is a very unlikely mark;
- In average, about 100 female SA freedivers must set their best mark ever so that a best individual record above $\approx$ 7 minutes (420 seconds) is observed;
- The maximum apnea time a female SA freediver can possibly set as her personal record is $\approx$ 10 minutes and 8 seconds (608 seconds), a much lower expectation than the over 16 minutes estimated by the Block Maxima approach.

We could also conclude that the right tail of the underlying distribution function to the freediving data at hand behaves in a way lighter than an Exponential-type tail, positioning itself in the Weibull max-domain of attraction.

4.1.1.3 Largest Yearly Observations Method

The inference performed in this section is based in the foundations laid in section 2.5 of Chapter 2 and section 3.1.3 of Chapter 3. But for the application of this Largest Observations method, we must look at our data in a different light.

In the two previous sections, in which we used the Gumbel and POT methodologies to infer on the data, we completely discarded one piece of information that we possess for each female freediver's personal best SA record – the year in which it was set.

For the LO method we must work with the k largest observations from each block, analogously to the BM method, but our data is not naturally divided into blocks. For the Gumbel approach,

we considered each individual freediver as a "block" with only one observation. This will not do now, for we need more observations for each block (and considering several records for the same freediver, as is for some available in the rankings from which our data was collected, would imply high correlation within each block). So we are now using the common data division in yearly blocks, being each new block formed by all the personal competitive best records set in that year, clearly implying that the blocks will not all have the same dimensionality. But since we will only use a small number of records set in each year, this will not pose a problem.

However, this new setup will not mean we are giving up our stationarity assumption just yet. We will use annual blocks, but still overlook if there is any evolution on the largest observations data through the years. Furthermore, this way of looking at the data is simply for the benefit of the method, hopping it will provide us with more accurate estimates for the **core parameters** of the Multivariate GEV model – we know these must be the same as the core parameters of the GEV distribution we have worked with so far.

Thusly, our further estimation of extreme value indicators of interest can process and be interpreted in the exact same way as before, being comparable with the previously obtained estimates.

Figure 4.33 (plotted by the code in Appendix A.19) shows how our sample of competitive female freediver's best personal records were set over the 13 years we considered, from 2002 to 2014. Note that only the 795 records set over 180 seconds have been represented, so we are dealing with the exact same sample that was the starting point for the two previous methodologies.

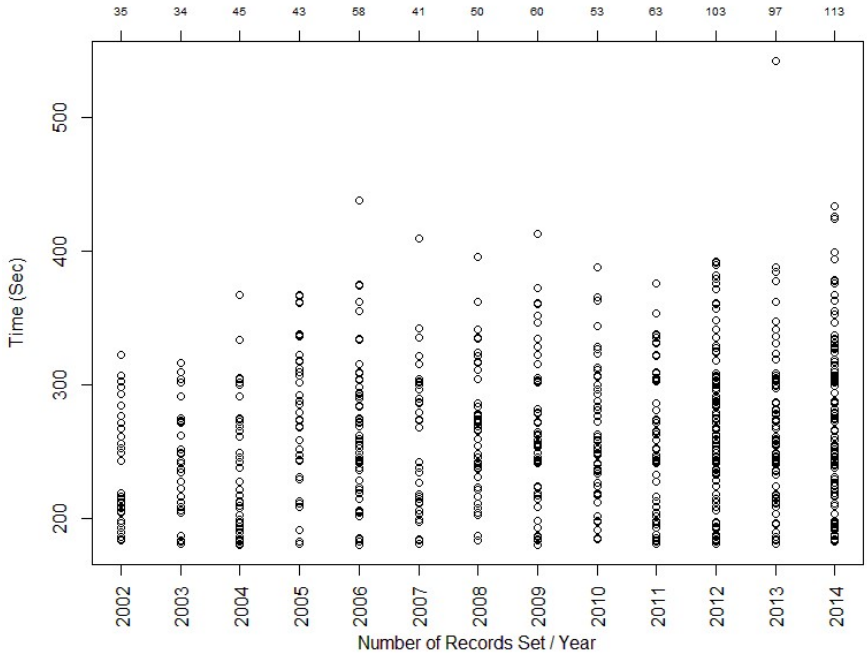

Figure 4.33: Competitive female SA freedivers' individual best records by year

The number of top observations usually considered from each block for the LO method is

around $k = 5$ or $k = 10$. Recall that considering $k = 1$ reduces the approach to the Gumbel method (not exactly as applied before, as here the blocks are yearly defined). We will do so here only for curiosity's sake, since considering this leaves us with a scarce sample of 13 observations, admittedly not enough for a trustworthy inference. As such we will pay little mind to those results. Also, we decided to further consider a higher level of $k = 20$ largest yearly observations, with the safeguard that such a large level includes smaller records, increasing the bias of the estimation. Hence, the four levels considered will be the $k = 1, 5, 10, 20$ largest yearly observations, and we plotted each corresponding block in Figure 4.34 (through the code in Appendix A.20).

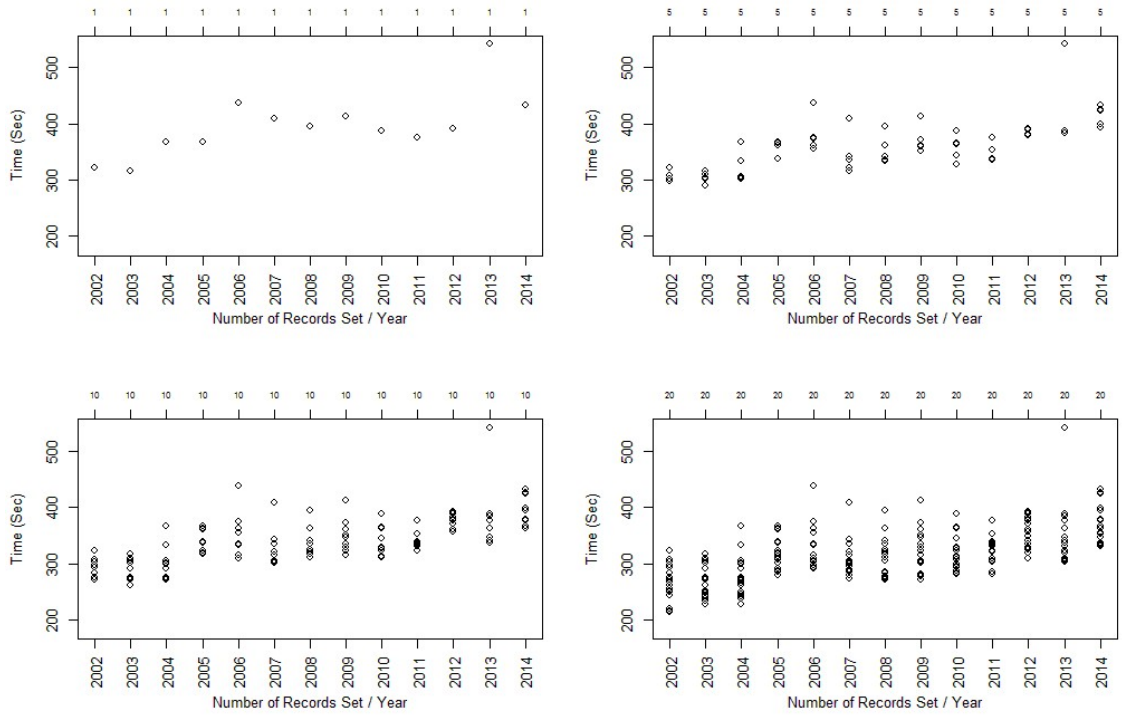

Figure 4.34: Largest 1 (top-left), 5 (top-right), 10 (bottom-left) and 20 (bottom-right) competitive female SA freedivers' individual best records by year

Table 4.15: Estimates for the Multivariate GEV fit to the largest k yearly competitive female SA freedivers' individual best records.

k	$logLikelihood$	$\hat{\xi}$ (se)	$\hat{\mu}$ (se)	$\hat{\sigma}$ (se)	$\widehat{P[X > 542]}$	$\widehat{\chi_{0.0001}}$	$\widehat{U(100)}$	$\widehat{x^F}$
1	-69.56247	-0.04630512 (0.1777963)	373.03330673 (13.8326061)	44.89364105 (9.7041644)	0.0158662	709.6494	559.0376	1342.551
5	-251.0371	-0.08664328 (0.07804513)	393.27184907 (8.62615659)	37.39337842 (4.45357910)	0.007593312	630.5433	535.1412	824.8504
10	-431.733	-0.1461297 (0.04308662)	401.3403458 (7.53612654)	35.5801950 (3.02701895)	0.002738205	581.4437	520.5074	644.824
20	-746.9938	-0.2104145 (0.02271049)	412.3416154 (6.57939140)	34.7841130 (1.73996901)	0.000681918	553.8499	514.8575	577.654

Table 4.15 shows, for each number of largest observations used from each block, the result of the Maximum Likelihood estimation of the parameters (ξ, μ, σ), the standard errors associated with these estimates, and the consequential estimation of the other indicators of interest: $P[X > 542]$, $\chi_{0.0001}$, $U(100)$ and, if applicable, x^F. The code can be found in Appendix A.21.

It is first evident that the estimate for the EVI is in all 4 cases negative, as we expected, and in line with the previously performed estimation. This allowed for the estimation of the finite right endpoint of the underlying short-tailed distribution. We can also see that as we increased the number of observations considered in the model, the estimate for the EVI became increasingly more negative – causing decreasing estimates for the other indicators such as the exceedance probability of the sample maximum and the right endpoint.

The case $k = 1$, as stated before, cannot be taken into much account, since the model if fitted to just 13 observations, which is not enough to guarantee a proper adjustment. Still, we can observe that the negative estimate of the EVI is rather close to 0, not as negative as we expect it should be. This indicates a not so light tail and produces the largest indicators' estimates seen so far in this dissertation. Particularly, it places the maximum apnea time possible at around 22 minutes and 22 seconds, much higher than our highest estimate so far of 16 minutes and 8 seconds, under the GEV maxima fit.

The case $k = 5$ can be analyzed with a little more care, since the fit is now made considering 65 observations, which still isn't much but should be enough for more accurate estimates. The EVI is estimated as twice as negative than in the case $k = 1$, and it is close to the estimates found under the GEV fit for the maxima sample. However, the location parameter is here placed much more to the right, having a higher value than before. This produces a right endpoint estimate of 13 minutes and 44 seconds, now smaller than the GEV fit's estimated 16 minutes.

When we retain $k = 10$ top observations from each year, working with a total of 130 observations, the EVI is estimated in a very similar value to the obtained from the GP fit in the POT approach. However, the location parameter is estimated to be very high (around 400 seconds) which influences the higher estimation of the other indicators of interest, when compared with the mentioned GP fit. However, the difference between the estimated right endpoints from this case (10 minutes and 44 seconds) and from the GP fit is less than 40 seconds, leading us to think this particular fitting is rather concordant with the GP fit, giving it some validity. Under this multidimensional fit, it is estimated a 0.2% probability of exceedance of the 542 seconds maximum.

Modeling the top $k = 20$ observations from each year means working with 260 observations. As stated, this might be too eager a choice, since it is including in the fit observations of lower value. It produces a very low EVI estimate of approx. -0.2, meaning a quite short tailed underlying d.f. F for our r.v. X. The corresponding estimate for the right endpoint is around 9 minutes and 37 seconds, a mere 35 seconds higher than the current world record, which does not seem likely to be true. Finally, we observe that the standard errors for the estimates of the core parameters are constantly decreasing with the increase in the number of records used. This illustrates the bias-variance tradeoff in choosing the number of top observations used, since, for instance, the estimates for the case $k = 20$ have the lowest se's but we believe are very biased.

4.1.2 Semi-Parametric Approach

We will now proceed to the semi-parametric analysis of our competitive female freedivers' data, based on the methodologies presented in section 3.2. Recall that, unlike the parametric inference performed so far, there is now no assumption on the underlying distribution to the data. We simply need to assume that $F \in \mathcal{D}_{\mathcal{M}}(G_\xi)$ for some real value of the EVI ξ.

Again, the r.v. X at hand represents the individual best record of a competitive female freediver in the discipline of Static Apnea, measured in seconds and with underlying d.f. F. The observed sample $(x_1, \ldots, x_n)$ has $n = 795$ elements (above 180 seconds). The tail sample fraction is denoted k, theoretically corresponding to an intermediate sequence of integers in the sense of (3.45).

Our focus here is in determining the suitable max-domain of attraction for our underlying distribution, with emphasis on the estimation of the determinant parameter ξ – the EVI – and with that information infer on indicators of interest regarding the data at hand (the same indicators already estimated from a parametric point of view, with which these new results will be compared). There is a special interest in the estimation of the right endpoint of the underlying distribution (if we conclude, as before, that it most likely belongs to the Weibull domain).

Taking all the parametric inference in consideration, we choose to discard from the beginning of this section the possibility that we could be dealing with a heavy tail, i.e., an underlying distribution belonging to the Frechét domain of attraction.

Statistical Testing for the Extreme Value Index Sign It was emphasized in section 3.2.1 the relevance of knowing the appropriate domain of attraction for our data, that is the sign of ξ, prior to the estimation of parameters, since it can help us select the most useful semi-parametric estimators and avoid meaningless estimation.

The parametric estimation procedures left us with two possible candidates for the domain of attraction to which F belongs: the Gumbel and the Weibull max-domains, the latter being the most likely to be the true domain of attraction of the distribution underlying to our sample. As such, we must employ the testing procedures presented for this setup with the objective of discerning if $\xi = 0$ or $\xi < 0$. Preference in the null hypothesis is given to the borderline Gumbel domain case – $H_0 : \ F \in \mathcal{D}_{\mathcal{M}}(G_0)$ –, here tested against two possible alternative hypothesis: $H_1^{(1)} : \ F \in \mathcal{D}_{\mathcal{M}}(G_\xi)_{\xi \neq 0}$, corresponding to the hypothesis in (3.51), and $H_1^{(2)} : \ F \in \mathcal{D}_{\mathcal{M}}(G_\xi)_{\xi<0}$, corresponding to the first set of hypothesis in (3.52).

Figure 4.35 shows the sample paths of the Ratio $R_n^*(k)$, Hasofer-Wang $W_n^*(k)$ and Greenwood $Gr_n^*(k)$ test statistics defined in (3.55), (3.57) and (3.59), respectively, versus the random threshold k, and the corresponding Gumbel (for the Ratio test) and Normal (for the Hasofer-Wang and Greenwood tests) quantiles that determine the rejection at the level $\alpha = 5\%$ of H_0 in favor of the two-sided alternative $H_1^{(1)}$. Figure 4.36 shows the same sample paths, but now the quantiles represented determine the rejection at the level $\alpha = 5\%$ of H_0 in favor of the Weibull alternative $H_1^{(2)}$. For details on the plotting of this Figures see Appendix A.22.

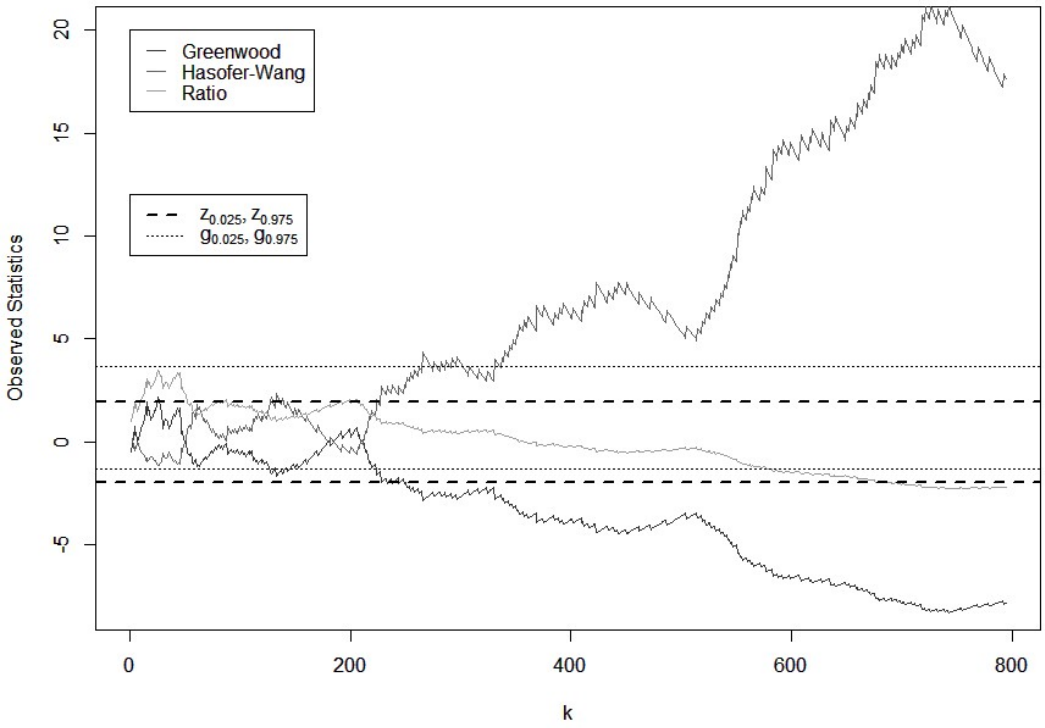

Figure 4.35: Sample paths for the Ratio, Hasofer-Wang and Greenwood test statistics and rejection regions for the two-sided alternative test

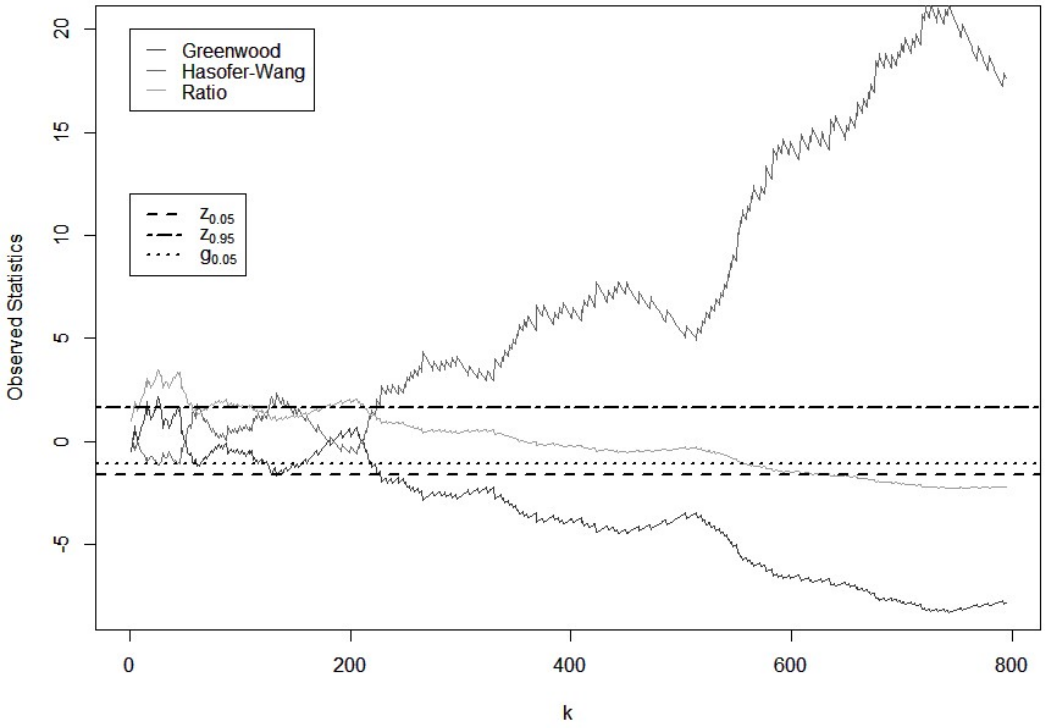

Figure 4.36: Sample paths for the Ratio, Hasofer-Wang and Greenwood test statistics and rejection regions for the Weibull alternative test

The Ratio test statistic is known to be the most conservative of the three here presented. We can see in both plots that this statistic points towards the non-rejection of the Gumbel domain hypothesis at the level 5% for most values of k, only indicating the rejection of said null hypothesis (both in favor of the two-sided and the one-sided Weibull domain alternatives) from values of k close to 600, which is admittedly too high of a random threshold for a total of 795 observations. However, both the Hasofer-Wang and Greenwood statistics' sample paths indicate the rejection of the Gumbel domain null hypothesis at the asymptotic level 5% in favor of both the two-sided and the one-sided Weibull domain alternatives for all values of k higher than the low 200's. This decision is made sooner for the Weibull domain alterative test than for the two-sided alternative one, and the first test statistic to reject H_0 is the Hasofer-Wang's, since it is the most powerful of the three for detecting the Weibull max-domain.

Since values of k in the low 200's can be considered appropriate random thresholds for a 795 element sample, it seems plausible to conclude from this analysis that the d.f. underlying to our data belongs to the Weibull max-domain. This comes in line with the conclusions from the parametric inference performed.

We presented in section 3.2.1 two further testing procedures for choosing the domain of attraction in a semi-parametric approach, the first based on the estimation of the right endpoint independently from the estimation of the EVI, and a second test for the finiteness of the right endpoint itself.

For applying the test for the Gumbel domain defined in (3.61), it is first necessary to compute the general right endpoint estimates as presented in equation (3.91). This is possible at this stage because it is independent from the estimation of the EVI itself. Note that this estimator, for each sample fraction k selected, utilizes half that number of top order statistics for computing the estimate of the right endpoint. Figure 4.37 shows the sample path of the general right endpoint estimator (constructed with the code in Appendix A.23), which is, as stated, always higher than the sample maximum, also marked in the plot.

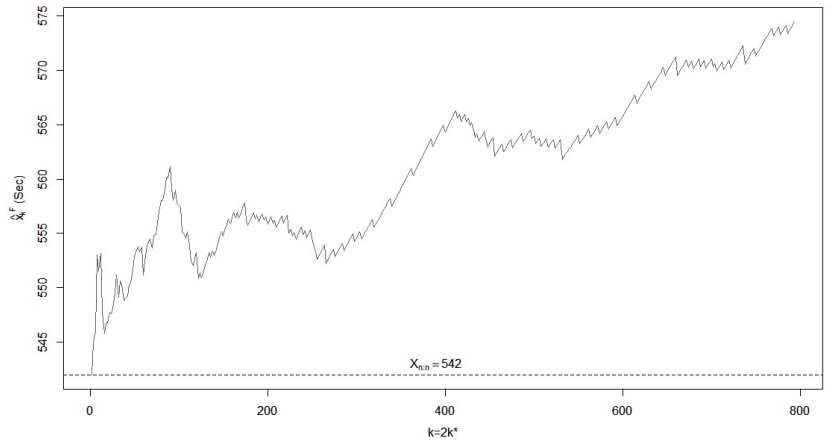

Figure 4.37: Sample path for the General Right Endpoint estimator for the competitive female freediver's personal best records data

Based on this estimation of the right endpoint, that will be discussed ahead, we drew the sample path of the $G^*_{n,k}(0)$ statistic versus the number of observations selected k. It can be seen in Figure 4.38, where also figure the Gumbel quantiles that define the rejection regions at the asymptotic level 0.05 for both the two-sided and one-sided Weibull alternative tests (see Appendix A.24).

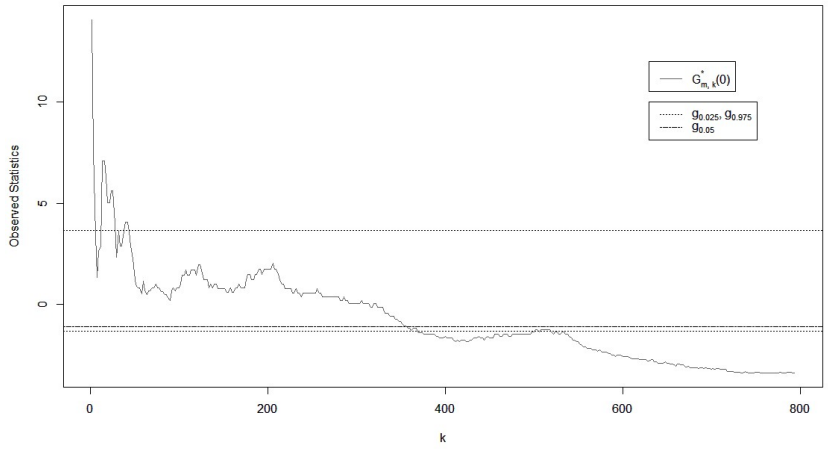

Figure 4.38: Sample path for the $G^*_{n,k}(0)$ statistic and rejection regions for the two-sided alternative and Weibull alternative tests for the competitive female freediver's personal best records data

According to this test statistic, the rejection of the Gumbel domain null hypothesis occurs only when almost half or more of the sample is used, that is, for values of k close to 400. This is now a considerably low threshold for a sample of 795 elements. This rejection occurs definitively and a little sooner for the one-sided alternative test than for the two-sided one, for which there is even an interval region of k around 500 that leads again to the non-rejection of the Gumbel domain hypothesis.

These results leave us with some doubts about the previous conclusion that a Weibull domain would be the correct domain to be elected, in detriment of the Gumbel domain. However, it is referred in Fraga Alves et al. (2016) that this test is more conservative than the Greenwood test, for example, for models with the EVI $\xi = -0.1$, performing similarly for models with $\xi = -0.2$. Since our parametric estimates of the EVI place it mostly in between these values, we will conclude the $G^*_{n,k}(0)$ statistic is in this case more conservative than the Greenwood statistic.

Finally, we will test for the finiteness of the right endpoint of the underlying distribution to our data, that allows us to determine if estimating the right endpoint is or is not a meaningless task, even if we are unable to elect with certainty the correct domain of attraction, between the Gumbel and Weibull candidate domains. The sample path of the T^*_1 statistic, defined in equation (3.66), can be seen in Figure 4.39 , along with the Normal distribution quantiles for the rejection at the asymptotic level of 0.05 of the null hypothesis in (3.62), i.e., the hypothesis that the right tail of F is unbounded (see Appendix A.25).

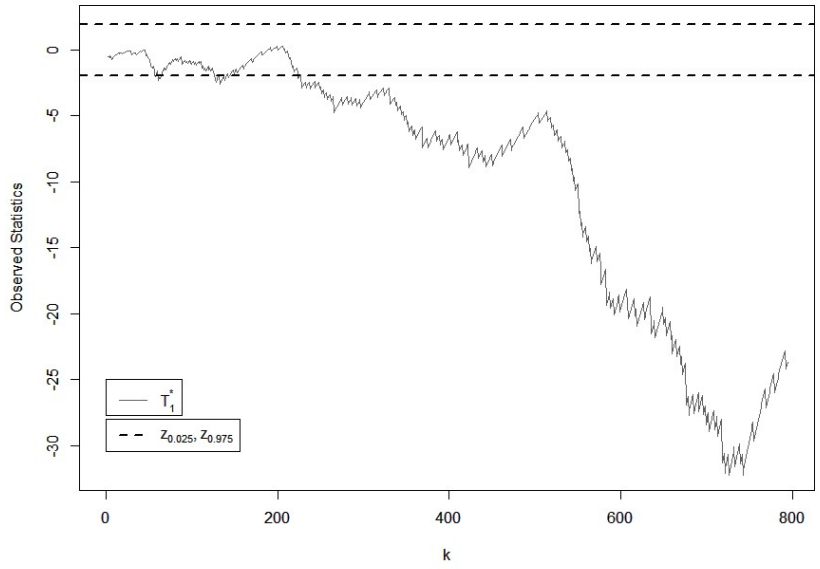

Figure 4.39: Sample path for the T_1^* statistic and rejection regions for the $x^F = \infty$ hypothesis for the competitive female freediver's personal best records data

Similarly to what was verified for the Hasofer-Wang and Greenwood statistics for testing the Gumbel max-domain, the rejection of the null hypothesis of a infinite right endpoint is pointed by this statistic above the random threshold given by k around 225 (low 200's). We are therefore confident in affirming that the underlying distribution to our sample has a finite right endpoint, as suggested by the parametric inference performed previously.

From this analysis we can once again conclude on the exclusion of the Fréchet domain of attraction possibility, and believe we are dealing with a short bounded-tailed distribution, most probably (but not surely) on the Weibull max-domain. This is to be taken into attention in the estimation to be performed ahead.

Estimation of the EVI and Choice of the Tail Sample Fraction As explained in section 3.2.5, in order to obtain valid and accurate estimates in a semi-parametric context we need to be working with an appropriate random threshold, defined by the tail sample fraction k selected, which should theoretically obey to some pre-set conditions. It was also seen that the practical choice of this value k is not simple, and an heuristic procedure was suggested to that end.

Since the semi-parametric estimators, analogously to the test statistics presented in the previous section, depend on the number of order statistics used in its computation, i.e., on k, we will start by selecting the appropriate estimators of the EVI for the setup at hand (given the results of the statistical tests presented) which will be used in the heuristic procedure for the choice of the "optimal" k, and calculate the corresponding estimates for all possible thresholds. This allows us to draw the sample path of each estimator, analyzing its behaviour and comparing it to the other estimators considered.

It is important to refer that we are assuming satisfied the extra specific conditions that are necessary for the validity of some of the estimators' properties, which were not verified neither specified in this dissertation.

We have discarded the possibility of a heavy tailed distribution underlying our sample, that is, the case $\xi > 0$ is out of play. This allows us to reject the Hill estimator for the EVI from the start, for it is only applicable in that case. We could choose to estimate the tail index with the Negative Hill estimator, since it is applicable when $\xi < 0$, which we suspect is the case for our data. However, this estimator is intended to be used in very short tails, with $\xi < -0.5$, and according to the preliminary estimation performed previously, we place our EVI always above -0.2. Thus, we reject this estimator and opt by the Generalized Hill estimator, including all the $\xi \in \mathbb{R}$ range.

Starting with the first semi-parametric EVI estimator introduced, the Pickands estimator, whose sample path can be seen in Figure 4.40 (corresponding code in AppendixA.26), we see a high variability throughout the plot typical of this estimator. Ignoring the first estimates that correspond to very small values of k (very high threshold and a small number of o.s.'s used), the variability of this estimator is still very visible, for example for values of k between 200 and 250, where the estimates go quickly from quite positive to quite negative values, and around $k = 400$, where the estimates drop considerably close to -1. For comparison purposes, we included in the plot the ML estimate of the EVI under the POT approach for the $u = 240$ threshold. Curiously, the Pickands estimates are close to the -0.16 POT estimate for values of k above around 225, roughly the same level above which we rejected the Gumbel domain hypothesis with the Greenwood and Hasofer-Wang test statistics.

The variability of the estimator is even more evident if we take into account the scale of the plot in Figure 4.40. A close up look for the $-1 < \hat{\xi} < 1$ interval shows very clearly the high instability of the sample path.

Note that the POT threshold considered of $u = 240$ implies the use of the 515 observations above this level, as shown before. This corresponds to $k + 1 = 515$ which means the "random" threshold considered for the POT approach was $X_{(514)}$, possibly a too low threshold. In this region, the Pickands estimates are significantly lower than $\hat{\xi} = -0.16$.

If we concentrate on the functional form of this estimator, we note it uses only 3 o.s.'s defined by the threshold provided. As such, it was computed for thresholds separated by lags of 4, leaving us with a quarter of the estimates other estimators that are computed for every value of k will have. This will cause some distortion in the application of the heuristic procedure for the choice of the optimal k – we will see ahead how to deal with this issue. Furthermore, its high variance can also bring some distortion to the measurements of the square distance between the estimates on which the heuristic is based.

As said before, we will not consider the Hill or Negative Hill estimators, so the next estimator analyzed is the Generalized Hill estimator, whose sample path can be seen in Figure 4.41 (see Appendix A.27).

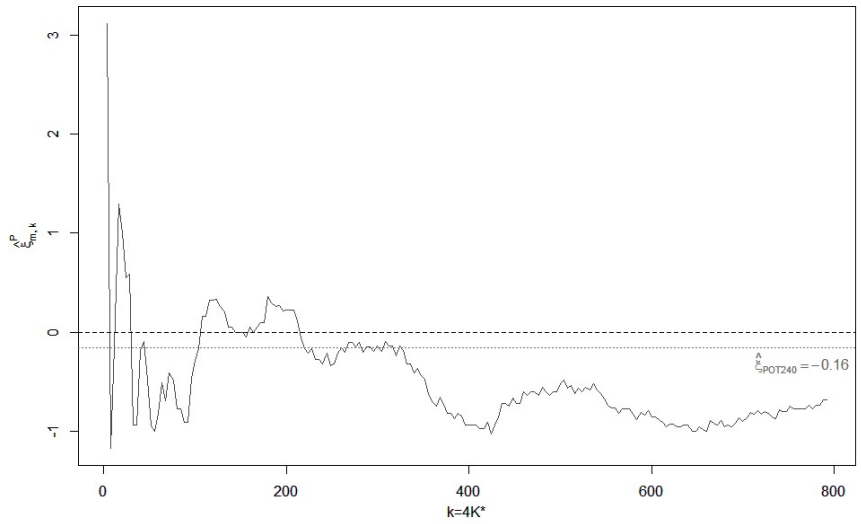

Figure 4.40: Sample path for the Pickands estimator of the EVI for the competitive female freediver's personal best records data

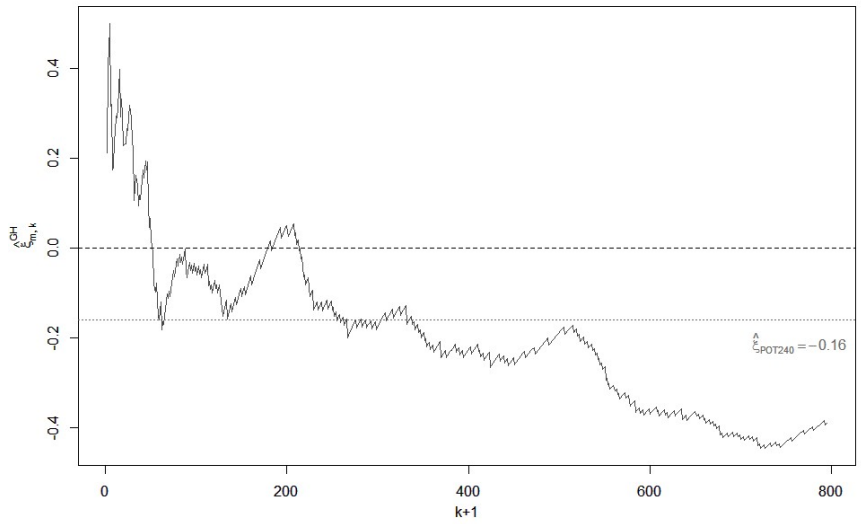

Figure 4.41: Sample path for the Generalized Hill estimator of the EVI for the competitive female freediver's personal best records data

At first glance the Generalized Hill estimator seems to be even more variable than the Pickands estimator, however if we attend to the scale of the plot, clearly that is not the case. This estimator, in addition to being negative for values of k above the low 200's, shows an approximately stable region roughly between $k = 220$ and $k = 330$ where the estimates are close to the POT over $u = 240$ ML estimate. Actually, around the threshold level of the POT approach, that we saw corresponded to $k = 514$, the Generalized Hill estimate is again fairly close to the POT estimate for the EVI.

The plot in Figure 4.42 shows the sample paths of both the Moment and Negative Moment estimators, resp. for the cases $\xi \in \mathbb{R}$ and $\xi < 0$. As we can see the paths are very similar, with the Negative Moment estimator dislocated down from the Moment estimator for the majority of the values of k, which is not surprising given the functional expressions of these estimators. Still, both provide mostly negative estimates for the EVI, and analogously to what was verified for the Generalized Hill estimator, there is a roughly stable estimation region between approx. $k = 220$ and $k = 330$. In this region, the Moment estimator is close to the POT approach ML estimate of $\hat{\xi} = -0.16$, with the Negative Moment estimator providing with considerably lower estimates (around -0.4). Recall these estimators are not invariant to changes in location. The corresponding code for the computation of this estimators can be found in Appendix A.28.

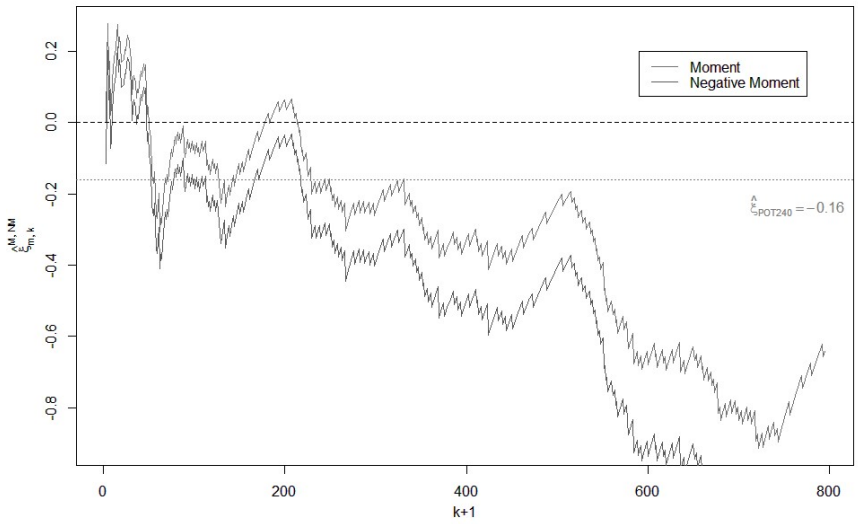

Figure 4.42: Sample paths for the Moment and Negative Moment estimators of the EVI for the competitive female freediver's personal best records data

Another non-shift invariant estimator for the case $\xi \in \mathbb{R}$ is the Mixed Moment estimator, whose sample path is shown in Figure 4.43 (see Appendix A.29). This is more satisfying than the previously considered estimators, since it has a clearly lower variability and points towards a estimates of the EVI not as low (not as negative) as the other estimators, throughout the path. This means the underlying distribution is expected to have a short tail, but not as short as expected based on the previous estimators. Moreover, this path crosses the POT approach with $u = 240$ ML estimate several times and keeps very close to that value through a significantly large interval of values of k, including the region between approx. $k = 220$ and $k = 330$, already referred as a stability zone for the previous estimators.

We begin to realize that the optimal value of k can possibly be around this stability region, where so far the estimators seem to be most concordant. However, this is simply a subjective impression given by the rough appreciation of the plots shown so far.

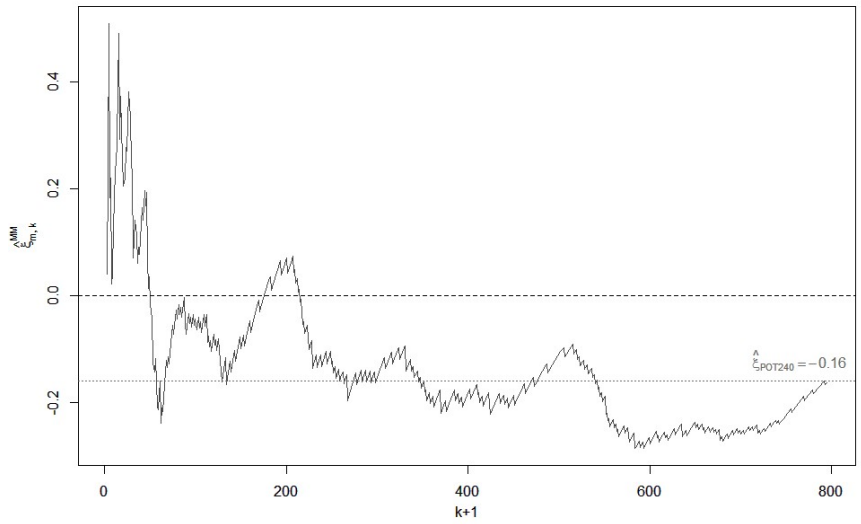

Figure 4.43: Sample path for the Mixed Moment estimator of the EVI for the competitive female freediver's personal best records data

Let us analyze how does the sample path for the Location Invariant Moment estimator behave, having the appealing quality of invariance of shift. The trajectory can be seen in the left plot of Figure 4.44, being the output of the code in Appendix A.30. The plot in the right shows the previously presented sample path for the Moment estimator, for comparison purposes.

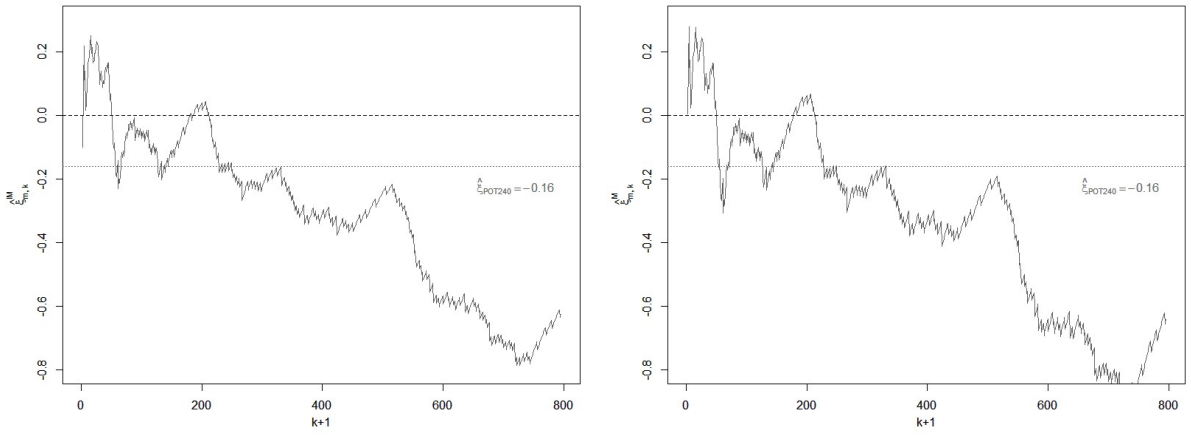

Figure 4.44: Sample paths for the Location Invariant Moment (left) and Moment (right) estimators of the EVI for the competitive female freediver's personal best records data

The paths for both estimators look very similar, and in fact the estimates obtained are roughly the same, except for the peak around $k = 500$ where the Location Invariant estimates fall below the Moment estimates, and for values of k larger than $\approx$550 – in that region, the Location Invariant estimator produces higher estimates than the original Moment estimator, but this region is not very interesting since it represents very low thresholds, that is, almost the

complete sample is used in the computation of the estimates, so the optimal k is unlikely to be found there. So this estimator gives us conclusions very similar to the ones drawn from the Moment estimator.

The other location and scale invariant alternatives for the Moment and Mixed Moment estimators presented were the respective PORT estimators. However, for using this estimators, there is the necessity of choosing the tuning parameter q. It is important to acknowledge that larger choices of this parameter reduce greatly the size of the sample we are working on. As such, we plotted the sample paths of the PORT-Moment and PORT-Mixed Moment estimators for a selection of values for the tuning parameter, before choosing the one to be considered in the heuristic procedure of choice of k.

The PORT-Moment estimators were calculated for the values $q = 0, 0.1, 0.2, 0.5$, and the corresponding sample paths can be seen in Figure 4.45 (output of the code in Appendix A.31).

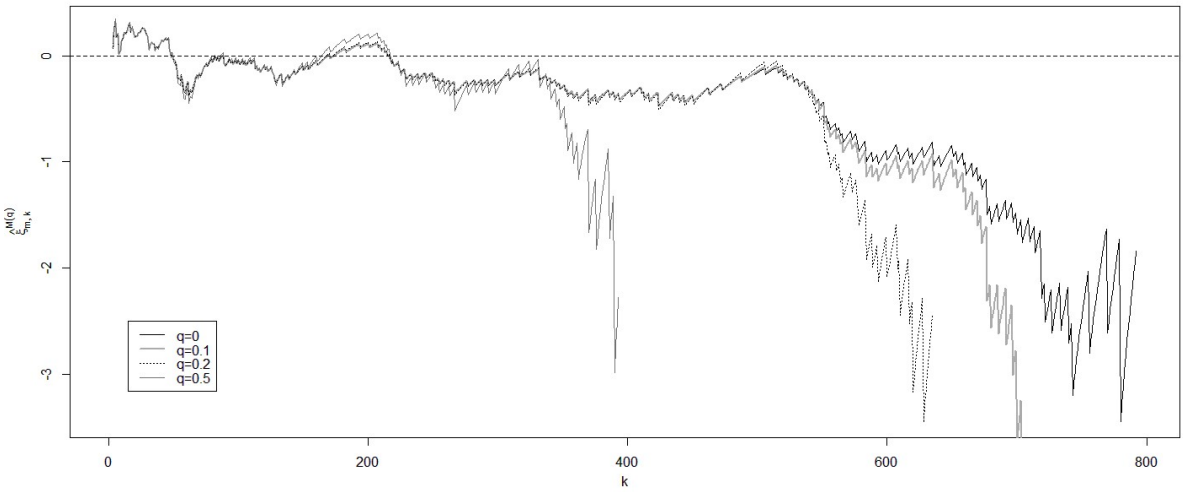

Figure 4.45: Sample paths for the PORT-Moment $q = 0, 0.1, 0.2, 0.5$ estimators of the EVI for the competitive female freediver's personal best records data

It is evident that considering a tuning parameter of $q = 0.5$ reduces the sample size by half, which is not desirable, and this estimator is from the four the one which most deviates from the rest for the values of k inferior to ≈350. Moreover, the other three estimators, that is considering $q = 0, 0.1, 0.2$, have very similar and even roughly regular paths up to values of k around 520. They all suffer an abrupt drop for the last thresholds k they consider, but as said before these regions do not have great importance (correspond to using almost the complete sample in the computations). Following the recommendation in Fraga Alves et al. (2009b), we will consider the estimator corresponding to the tuning parameter $q = 0.1$ for the rest of this semi-parametric analysis.

The PORT-Mixed Moment estimators were calculated for the values $q = 0, 0.01, 0.1, 0.2$, and the corresponding sample paths can be seen in Figure 4.46 (output of the code in Appendix A.32).

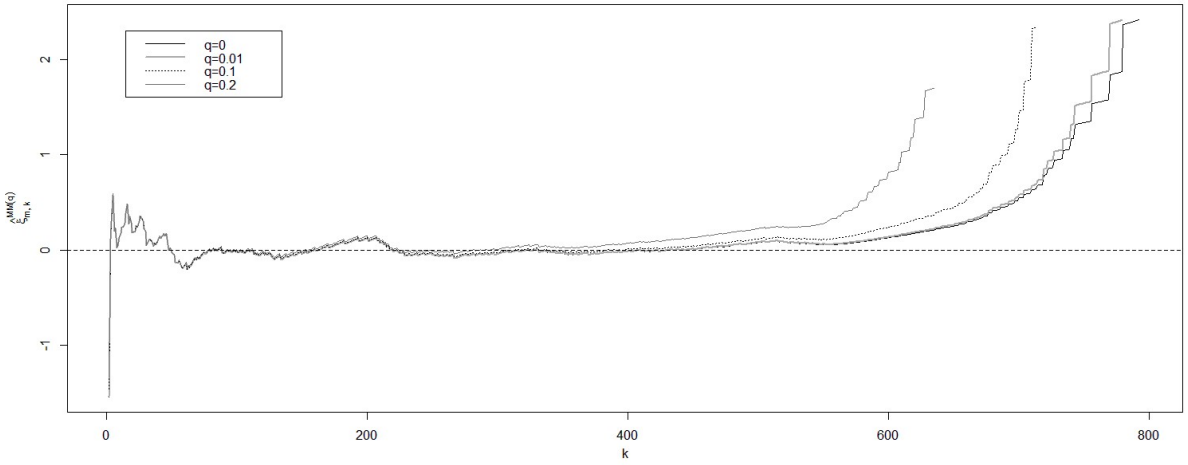

Figure 4.46: Sample paths for the PORT-Mixed Moment $q = 0, 0.01, 0.1, 0.2$ estimators of the EVI for the competitive female freediver's personal best records data

These samples paths are very distinct from all the presented above, and very similar between themselves. Unlike the estimators analyzed so far whose sample paths where mostly on the negative part of the plot, these PORT-Mixed Moment estimates concentrate around the 0 line, varying between positive and negative estimates for the EVI. Furthermore, for values of k above $\approx$400, the estimates are exclusively positive and increasing. This abnormal behaviour leads us to discarding any PORT-Mixed Moment estimator form the rest of the semi-parametric analysis, in particular from the heuristic procedure for the choice of the optimal tail sample fraction. However, we would still like to note that what was considered a stability region for some of the previously shown estimators is in this case the region where the three PORT-MM estimators with $q = 0, 0.01, 0.1$ look too the most stable and present (slightly) negative estimates for the EVI.

Figure 4.47 comprises the sample paths of all the considered estimators above (with the exception of the PORT-Mixed Moment estimators). In this plot it becomes even more clear the higher variability of the Pickands estimator when compared with the rest of the estimators. As we have been realizing through this analysis, the region in which the most estimators seem concordant is in fact the "stability region" around $k = 220$ – ignoring, of course, the high volatility area of the very low values of k. We then expect that using these estimators in the heuristic procedure proposed by Henriques-Rodrigues et al. (2011) we will obtain an "optimal" sample fraction in this region. For the code that results in this image see Appendix A.33.

Recall the definition of the heuristic procedure in equation (3.96):

$$k^{opt} = \arg\min_{k} \sum_{(E,J)\in\mathbb{E}:E\neq J} \left(\hat{\xi}^{E}_{n,k} - \hat{\xi}^{J}_{n,k}\right)^2,$$

where the set of estimators for this case study is $\mathbb{E} = P, GH, M, NM, MM, IM, M(0.1)$.

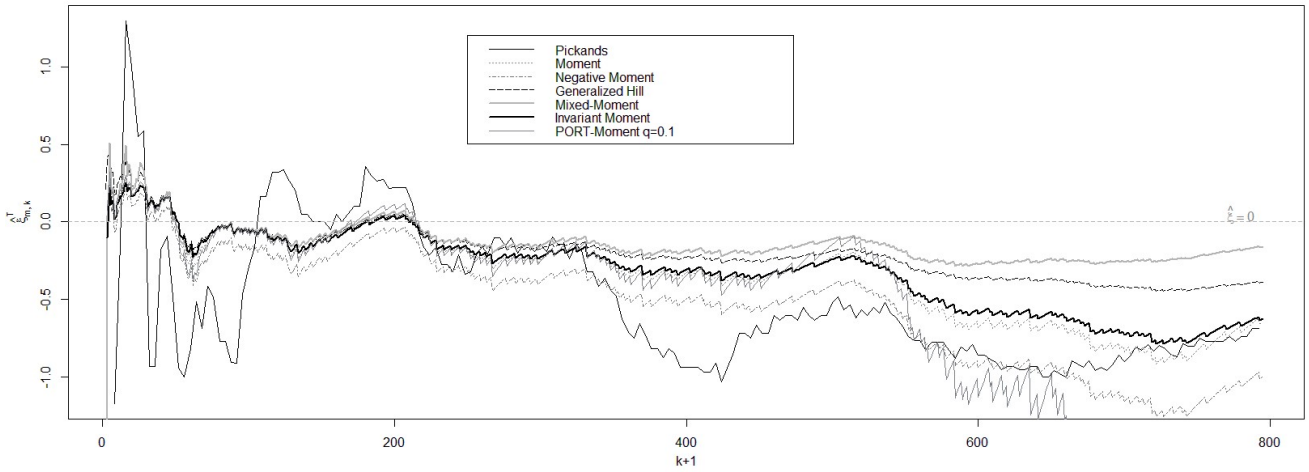

Figure 4.47: Sample paths for the Pickands, Generalized Hill, Moment, Negative Moment, Mixed Moment, Location Invariant Moment and PORT-Moment $q = 0.1$ estimators of the EVI for the competitive female freediver's personal best records data

We mentioned before a difficulty in the application of this procedure with the inclusion of the Pickands estimator. Because of the way this estimator is computed, it will only be comparable for values of k multiple of 4, and as such we have to apply the heuristic only for the estimates from each estimator corresponding to values of k multiple of 4. This will still cover the scope of k we wish to analyze, and we expect it to not be a very significant restriction, since the variation in the estimates is not commonly very abrupt for very close values of the tail sample fraction. An alternative course of action would be to discard the Pickands estimator, but we chose here not to because it helps the heuristic not to chose very small values of k, where the other estimators are very close, but we know the estimates are not reliable. This derives from the enormous variability of the Pickands estimator, that comes in handy at this stage.

The "optimal" tail sample fraction we obtained form this procedure is $k^{opt} = 216$, which means the estimators are calculated from the 217 top observations from the sample, corresponding to the random threshold $x_{(217)} = x_{(795-217+1):795} = x_{(579):795} = 295$ (higher than the $u = 240$ seconds used for the POT parametric approach and closer to the alternative threshold considered $u = 300$ seconds). This results from the code in Appendix A.34, which translates the above detailed heuristic. It is also possible to plot the sample path of the distance measure on which the heuristic is based, as can be seen in Figure 4.48.

The result of $k^{opt} = 216$ is not surprising, since it falls close to the region we had observed the most estimators were concordant. Moreover, we can see in Figure 4.48 that what we called the "stability region" for some of the estimators comes translated also in the heuristic's plot. Between k values of $\approx$200 and $\approx$350, the measure of distance used has very low values throughout, and there is very little variability (the existing variation in this measure, despite being small, is masked by the scale of the plot).

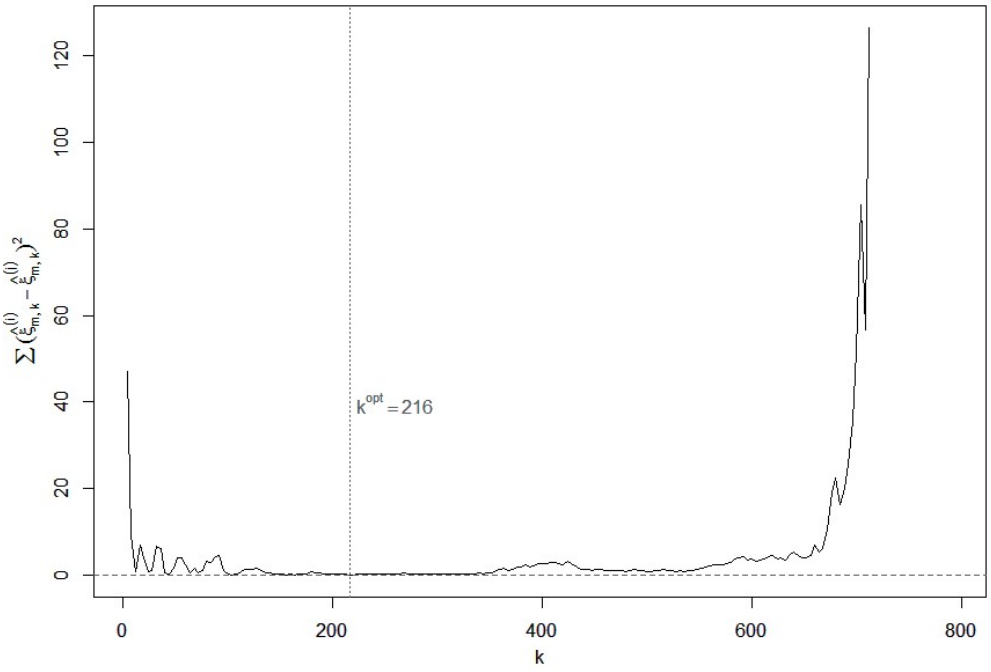

Figure 4.48: Sample pats for the distance $\sum_{(E,J)\in\mathbb{E}:E\neq J}\left(\hat{\xi}^{E}_{n,k}-\hat{\xi}^{J}_{n,k}\right)^2$ of the EVI's estimators for the competitive female freediver's personal best records data

We have been comparing the sample paths of the semi-parametric estimators with the EVI estimate obtained by the Maximum Likelihood method for the POT approach, with the empirically chosen threshold $u = 240$ seconds. However, there is one more semi-parametric estimator we have still to consider: the one that derives from fitting a Generalized Pareto model through the ML method to our data considering all the possible random thresholds, as presented in section 3.2.2. Figure 4.49 shows the sample path of this POT-ML estimator, plotted against the sample paths of the Generalized Hill, Mixed Moment and Location Invariant Moment estimators (see Appendix A.35). Also marked in the plot is the optimal tail sample fraction obtained from the heuristic procedure $k^{opt} = 216$ and the corresponding EVI estimates for the considered estimators.

It is visible the concordance of this POT-ML estimator with the other three considered (which are fairly representative of the general behaviour of the estimators plotted in Figure 4.47) in the region where we found the optimal value of k, which is indicative of a good choice, since the POT-ML estimator was not considered in the heuristic procedure. We also find that the variability of this estimator is smaller than that of the others represented – in the final section of the plot, where the other estimators traditionally behave poorly, the POT-ML estimator is more stable (in the zone of the threshold chosen for the parametric approach of $u = 240$, corresponding to $k = 514$), even though with a slight decreasing tendency.

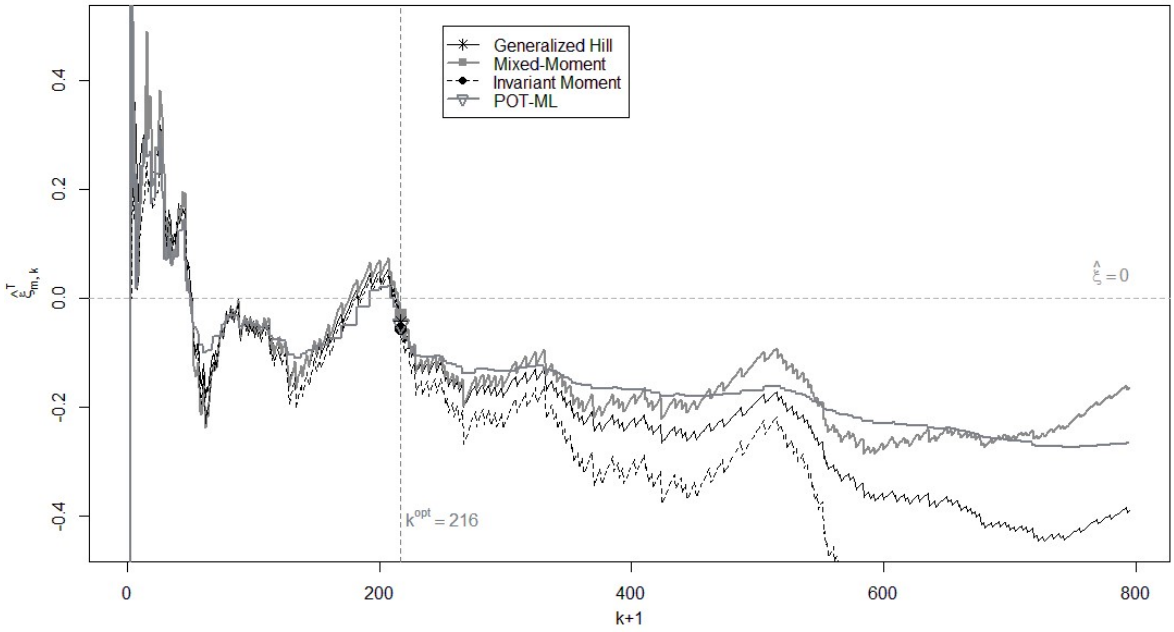

Figure 4.49: Sample paths for the POT-ML, Generalized Hill, Mixed Moment and Location Invariant Moment estimators of the EVI for the competitive female freediver's personal best records data

Now that we found the heuristically optimal tail sample fraction for the computation of the semi-parametric estimates, Table 4.16 shows the result of the selected estimators for the EVI estimation at $k^{opt} = 216$. Furthermore, knowing the form of the asymptotic variance of the estimators and assuming, as mentioned, the specific conditions that allow for this property to stand, we were also able to obtain 95% CI's for the tail index ξ for each estimator (with the previously declared exception of the Location Invariant Moment estimator). The computation of this point and interval estimations was performed by the code in Appendix A.36.

Table 4.16: EVI semi-parametric estimates and 95% CI's at $k^{opt} = 216$ for the competitive female freediver's personal best records data.

Estimator	$\hat{\xi}_{n,216}$	95% **CI**
Pickands	-0.05658353	(-0.5343481 ; 0.4211810)
Gen. Hill	-0.04258196	(-0.17072658 ; 0.08556267)
Moment	-0.0518952	(-0.18082569 ; 0.07703529)
Neg. Moment	-0.1625763	(-0.29349741 ; -0.03165516)
Mixed Moment	-0.02993382	(-0.1601786 ; 0.1003109)
Loc. Inv. Moment	-0.05632617	-
PORT-Moment	-0.03062056	(-0.16081277 ; 0.09957164)
POT-ML	-0.04773188	(-0.17472508 ; 0.07926132)

The first aspect we should note is that every point estimate for the EVI is, without exception, negative. However, apart from the Negative Moment estimate, which is more than 3 times lower than the others, the estimates are considerably close to 0. Looking back at the parametric GEV fit to the maxima sample, the ML estimate for the EVI was placed around -0.06, which is almost the same as the Pickands, Moment and Location Invariant Moment estimates of less than -0.05, and even sufficiently similar to the Generalized Hill and POT-ML estimates of about -0.04. The Mixed Moment and PORT-Mixed Moment estimates are the closest to 0, at approx. -0.03. These estimates are all of the same approximate magnitude of the GEV-ML estimate, and we can see the corresponding 95% CI's have a negative lower limit and a positive upper limit. This is coherent with our analysis of the Greenwood and Hasofer-Wang test statistics' sample paths, which showed that around the region in which we find our $k^{opt} = 216$ we went from not rejecting the Gumbel domain hypothesis to rejecting it in favour of the Weibull max-domain. Thus, it is understandable that the confidence intervals for the EVI computed at such a value of k include the possibility $\xi = 0$. The Negative Moment estimator is the dissonant case here, having much lower estimates for the EVI and a completely negative CI. In fact, its estimate at $k^{opt} = 216$ is around -0.16, the same as the obtained in the parametric POT approach with the threshold $u = 240$. However, we now suspect this threshold might not have been the most suitable choice.

This estimation results are pleasing in the sense they point towards a short-tailed distribution underlying our freedivers' data, with a finite right endpoint, as we expected.

Knowing the optimal tail sample fraction also allows for the estimation of the location and scale attraction coefficients as presented in section 3.2.3, which depend only on k and not directly on the EVI estimation. We then estimate the location attraction coefficient as $\hat{b}\left(\frac{n}{216}\right) = x_{(217)} = 295$ (seconds) and the scale attraction coefficient as $\hat{a}\left(\frac{n}{216}\right) = 37.95919$ (see Appendix A.37).

Estimation of Other Indicators of Interest In a similar fashion to what was made for the several parametric approaches, we will now use the EVI and attraction coefficients' semi-parametric estimates, obtained above for the tail sample fraction $k^{opt} = 216$, to compute semi-parametric estimates of the now familiar indicators of interest about the r.v. X of the freedivers' competitive records. However, we must now rely on the estimation methods and functional expressions presented on section 3.2.3 for finding $P[\widehat{X > 542}]$, $\widehat{\chi_{0.0001}}$, $\widehat{U(100)}$ and $\widehat{x^F}$, for each of the EVI's estimates obtained before.

This estimation was performed by running the code presented in Appendix A.38 and the corresponding results are shown in Table 4.17. Note here that the estimator of the scale attraction coefficient to be used in the estimations deriving from the Location Invariant Moment estimator is different from the presented before. According to Ferreira et al. (2003), we should here estimate $\hat{a}\left(\frac{n}{k}\right) = N_{n,k}^{(1)}\left(1 - \hat{\xi}_{n,k}^{IM}\right)$, which produces in this case $\hat{a}\left(\frac{n}{216}\right) = 38.15978$ (see Appendix A.37).

The abnormal case in this set of estimates corresponds to the Negative Moment estimator of the EVI: the estimate for the tail index is sufficiently negative so that it induces an estimate for the right endpoint inferior to the value of the sample maximum (highlighted in bold in Table 4.17), and consequently a null probability of exceedance of the current world record of 542 seconds.

Now excluding the Negative Moment estimates, we can see that all the exceedance probability $P[X > 542]$ estimates are located around 0.01% which is, if we look back at the parametric inference, the approximate estimate for this exceedance probability obtained under the GEV fit to the maxima sample. Furthermore, the estimators are concordant in placing $\widehat{U(100)}$ at about 410 seconds, but present more significant variations in estimating the extremal quantile $\widehat{\chi_{0.0001}}$, from the 536 seconds from the Pickands estimator to the 562 seconds from the Mixed Moment estimator (maximum variation of 26 seconds).

The most significant discrepancies between the various estimates are observed for $\widehat{x^F}$, which is perhaps the indicator we are here most interested in understanding. The Pickands estimator provides us with the lowest estimate of the right endpoint at 965 seconds, close to the 972 seconds from the Location Invariant Moment estimator – recall that the parametric GEV fit for the maxima sample provided the ML estimate of 968 seconds for x^F. The Mixed Moment estimator (which yields the highest EVI estimate, see Table 4.16) induced the largest estimate for the right endpoint, at 1563 seconds – 26 minutes and 3 seconds, the highest estimate obtained so far for both the parametric and semi-parametric approaches to this data set. This corresponds to an approximate 10 minute variation between semi-parametric estimates for the right endpoint, all computed for the tail sample fraction $k^{opt} = 216$. This is clearly very significant, and might be pointing to the existence of a different problem here: the optimal choice of k for the estimation of the EVI may not be the optimal choice when our aim is to estimate the right endpoint of the d.f. underlying to our data.

Table 4.17: Semi-parametric estimates for the indicators of interest at $k^{opt} = 216$ for the competitive female freediver's personal best records data.

Estimator	$\widehat{P[X > 542]}_{216}$	$\widehat{\chi_{0.0001}}_{216}$	$\widehat{U(100)}_{216}$	$\widehat{x^F}_{216}$
Pickands	8.125943e-05	536.9936	409.3317	965.8522
Gen. Hill	0.0001333336	549.8463	411.9318	1186.438
Moment	9.685374e-05	541.1956	410.1935	1026.459
Neg. Moment	0	463.9261	391.9921	**528.4854**
Mixed Moment	0.0001954289	562.2757	414.3496	1563.103
Loc. Inv. Moment	8.661546e-05	538.5017	414.2166	972.4788
PORT-Moment	0.0001916688	561.5799	414.2166	1534.663
POT-ML	0.0001122337	545.012	410.9662	1090.259

Before further dwelling on this issue, there is one more estimator for the right endpoint we have yet to consider, that has not been added to Table 4.17 because it does not depend on any estimation of ξ: the General Right Endpoint estimator, defined in eq. (3.91) in section 3.2.3, already applied in this case study for the construction of the a test statistic for the EVI sign. The sample path of this estimator has been shown in Figure 4.37 above (attend on the fact that on the horizontal axis is represented the tail sample fraction values, not the number of observations used in the computation). For the heuristically optimal tail sample fraction $k^{opt} = 216$ the corresponding general right endpoint estimate is around 556 seconds, a mere 14 seconds higher

than the current world record. This adds force to the conviction that $k^{opt} = 216$ might not be appropriate for the right endpoint estimation.

To find the appropriate tail sample fraction for the right endpoint estimation we can apply once more the heuristic procedure suggested by Henriques-Rodrigues et al. (2011) used for finding the previous $k^{opt} = 216$. The slight modification is that the distance at each value of k is measured between the several estimates for the endpoint, instead of those for the EVI. Moreover, in this case, we will not consider the Pickands derived estimates, since they have been shown to condition the points of k we assess. However, we will include the general right endpoint estimates, which are only computed for $k^* = k/2$, thus we can only asses the even values of k – we make this compromise because the general right endpoint's sample path shows promise of being a useful and appropriate estimator.

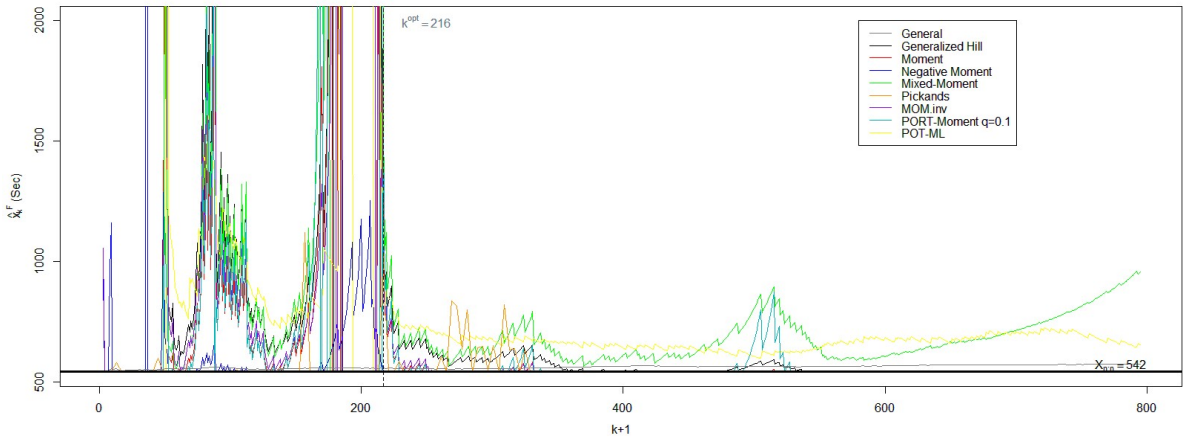

Figure 4.50: Complete sample paths for right endpoint estimators for the competitive female freediver's personal best records data

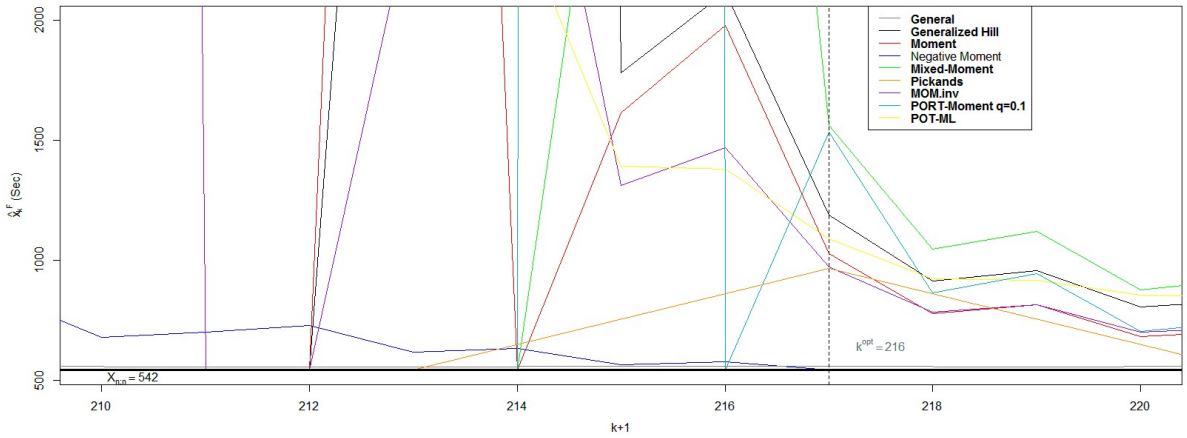

Figure 4.51: Sample paths for right endpoint estimators at $k \in [209; 219]$ for the competitive female freediver's personal best records data

Thus, for the reapplication of the heuristic we computed the sample paths of the endpoint estimators, plotted against the $k+1$ number of observations for the estimates. The code for the heuristic and for the plots of the paths in Figures 4.50 and 4.51 are in Appendix A.39.

Not many specific conclusions can be drawn from Figure 4.50, apart from observing the high variability of the estimators' trajectories, specially for values of k up to around 225. Furthermore, we see that from that point forward, most of the estimators are inadmissable, repeatedly falling below the 542. However, the general right endpoint estimator stands out as the less variable estimator, and we know that it is, by definition, always larger than the sample maxima.

In the close-up of this plot in Figure 4.51 we are able to clearly see that the $k^{opt} = 216$ chosen falls in a highly unstable zone for most of the estimators, even though the only inadmissable estimate is obtained from the Negative Moment estimator – all the others are, as seen in Table 4.17, larger than the sample maxima 542. The need for choosing a more appropriate tail sample fraction in thusly confirmed, but there is one possible problem to keep in attention.

As mentioned in section 3.2.5, this heuristic procedure for the right endpoint estimation can be problematic and must be applied with caution, due to the abundance of non-admissible estimates that tend to appear. We have observed this fact in the plotting of the sample paths, where there is a large region of values of k for which most of the estimates have the same value as the sample maxima, since otherwise they would have even lower and inadmissable values. This leads to many false k^{opt} resulting from the heuristic procedure. See, for instance, a section of the plot in Figure 4.50 between 0 and ≈ 50, where it is barely observable the sample path of the general right endpoint estimator, and no other estimators' trajectories are visible. Surely at this region the heuristic procedure will take some of its lowest values, but the corresponding sample fractions k are not of interest for the estimation.

So, in order to get a better understanding of what are the possible candidates for the optimal tail sample fraction in this case, we plotted in Figure 4.52 the sample path of the distance measure on which the heuristic is based, analogously to what had been done for its previous application.

As predicted, the unrestrained application of the heuristic procedure, as detailed in Appendix A.40, leads to the "optimal" tail sample fraction value of $k^{opt} = 2$, which is absurd. We see in the plot of the procedure that in the mentioned region of k between 0 and ≈ 50, where most of the estimators "disappeared" from the sample path plot in Figure 4.50, the squared distance between the estimates is close to 0. Thus, a more attentive analysis of this trajectory must be made for selecting appropriate intervals of k for applying the heuristic.

In the plot of Figure 4.52 are marked three peaks in regions which seemed reasonable at the time of the analysis, as well as the previously chosen $k^{opt} = 216$, clearly falling in a region where the heuristic function has a very high value of the squared distances between estimates. The three chosen peaks correspond to the tail sample fractions $k^{opt(1)} = 128$, $k^{opt(2)} = 266$ and $k^{opt(3)} = 370$. The sample paths of the estimators in Figure 4.50 were then plotted close-up in the regions of this values in Figures 4.53, 4.54 and 4.55, respectively (see Appendix A.41). The estimators in **bold** in the legends are the ones with admissible value at the k^{opt} in question.

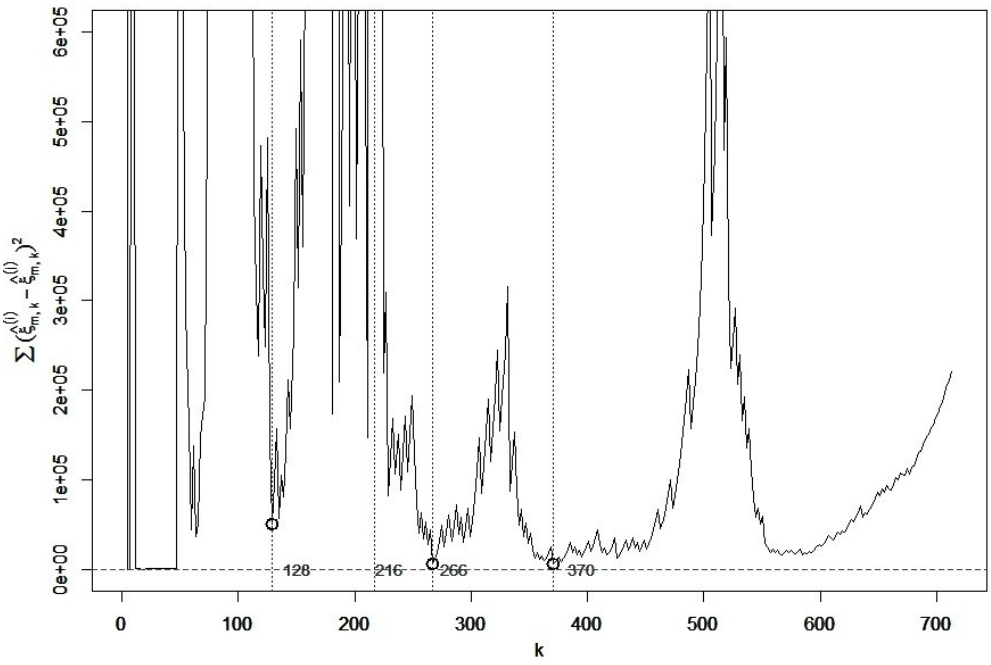

Figure 4.52: Sample path for the distance measure from the heuristic for the right endpoint estimators for the competitive female freediver's personal best records data

Table 4.18 summarized the results of the estimation of the right endpoint and of the EVI at the new values found for the tail sample fraction. Note that since the general estimator is independent from any EVI estimation, there is no associated estimate for ξ.

Table 4.18: Semi-parametric right endpoint (and EVI) estimates at $k^{opt(1)} = 128$, $k^{opt(2)} = 266$ and $k^{opt(3)} = 370$ for the competitive female freediver's personal best records data.

Estimator	$k^{opt(1)} = 128$		$k^{opt(2)} = 266$		$k^{opt(3)} = 370$	
	$\widehat{x^F}$	$\hat{\xi}$	$\widehat{x^F}$	$\hat{\xi}$	$\widehat{x^F}$	$\hat{\xi}$
Gen. Hill	626.3028	(-0.1506715)	563.9837	(-0.1987393)	542	(-0.2409232)
Moment	542	(-0.2251541)	542	(-0.304273)	542	(-0.3715246)
Neg. Moment	542	(-0.3392178)	542	(-0.44478)	542	(-0.5387207)
Mixed Moment	606.2336	(-0.1609138)	566.8956	(-0.1967292)	568.932	(-0.2149905)
Loc. Inv. Moment	551.8164	(-0.1931988)	542	(-0.2681138)	542	(-0.33682)
PORT-Moment	542	(-0.2485503)	542	(-0.3511495)	542	(-0.426722)
POT-ML	767.797	(-0.1040006)	692.1949	(-0.137072)	661.0832	(-0.165718)
General	551.4947	-	552.2374	-	561.323	-

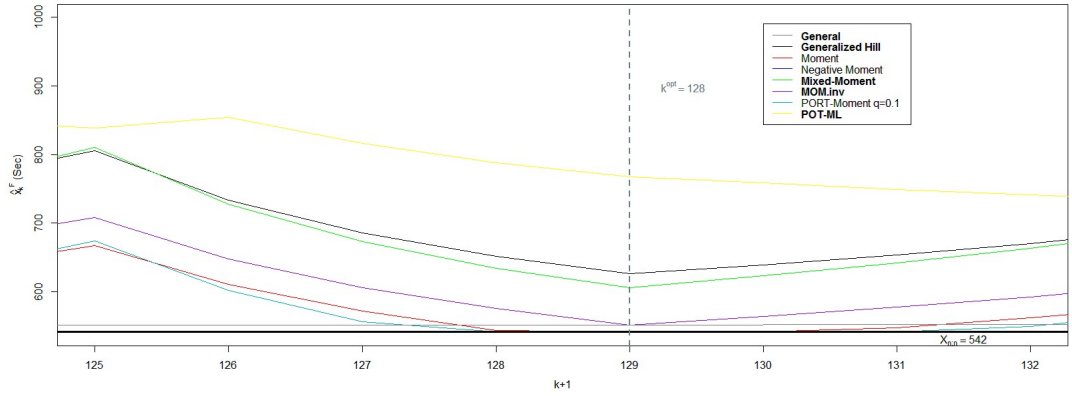

Figure 4.53: Sample paths for right endpoint estimators at $k \in [124; 131]$ for the competitive female freediver's personal best records data

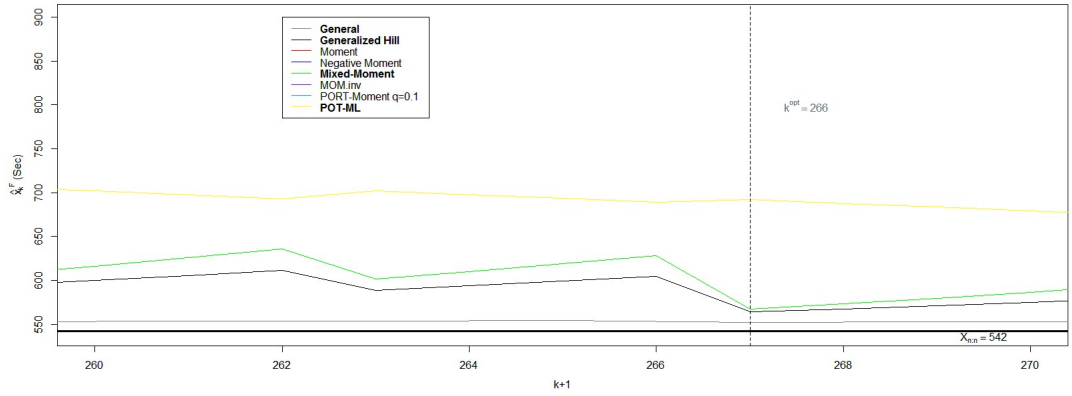

Figure 4.54: Sample paths for right endpoint estimators at $k \in [259; 269]$ for the competitive female freediver's personal best records data

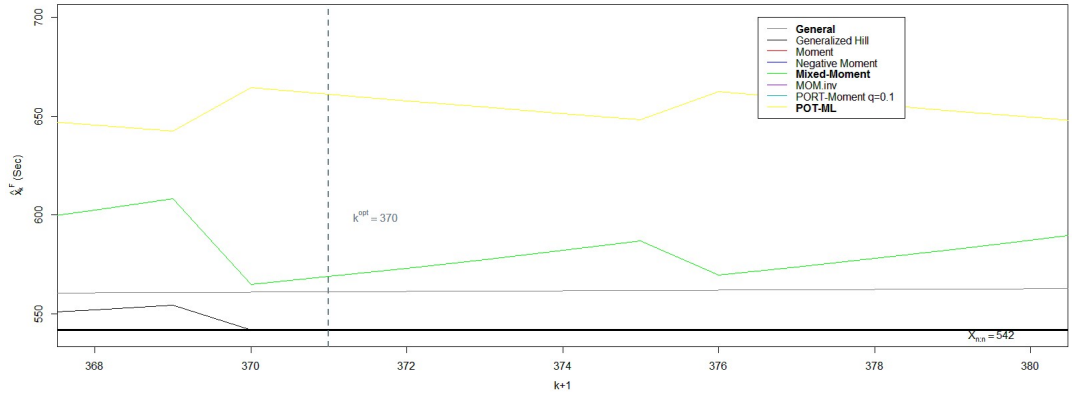

Figure 4.55: Sample paths for right endpoint estimators at $k \in [367; 379]$ for the competitive female freediver's personal best records data

First of all, note that when considering $k^{opt(1)} = 128$, only 5 out of the 8 estimators considered for the right endpoint have an admissible value (above 542), and this number drops to 4 when considering $k^{opt(2)} = 266$ and is even lower with just 3 out of 8 admissible estimates when the tail sample fraction is $k^{opt(3)} = 370$. These come accompanied by an overall decrease in the EVI estimates that appear too negative, implying underlying distributions with very short tails and consequently estimating the right endpoint lower than the sample maximum. The previously considered $k^{opt} = 216$ induced only one inadmissable estimate, but all the others were too different from each other, preventing us from drawing a definite conclusion about the true value of the endpoint.

The focus on these alternative choices of k has not provided us with much more useful information about our sample. It appears that the only safe conclusion to be drawn from this is that, in a broad view, the true right endpoint for the underlying distribution to freediving records seems to be lower than previously estimated, certainly not raising to the highest estimated limit of a 26 minute breath hold.

The semi-parametric estimation performed for the tail sample fraction $k^{opt} = 216$, comprised in Table 4.17, is fairly close to the results obtained in the parametric analysis, for the fitting of the GEV model with negative EVI of the sample of maxima, as mentioned before. This leads us to the conclusion that the GEV analysis performed, under the current state-of-the-art, is satisfyingly valid. The most dissonant indicator is the right endpoint, but the last approach made to this parameter in this section makes us believe that the true right endpoint must be located, at best, around the 16 minute mark (960 seconds), still an admirable value considerably higher than even the current male world record of 11 minutes and 35 seconds (and of course the female 9 minutes and 2 seconds world record).

4.2 Testing the Stationarity Assumption

Recall that for all the analysis and inference performed to this point in this dissertation, we worked under the stationarity assumption, that is, in a state-of-the-art setup, where we admitted that the passage of time does not have influence on the estimation results. It is easy to let our common sense judge this assumption as an overstep – surely, with the growth of the sport through the years and with more athletes joining the competition, we will be able to get more (in number) and more extreme observations of the best personal records. We will analyze from this point forward if and how the cofactor *time*, measured in years, changes the conclusions we obtained from the stationary parametric inference.

We will be referring to section 3.1.4 for the statistical setup, keeping in mind that the non-stationarity approach to extreme value data does not follow a common well-known modelling theory or procedure, as seen before, but mostly relies on a more pragmatic take depending on the type of non-stationarity present and on the experience and sensibility of the analyst. In this case study we will follow the methodology detailed in Coles (2001) and Fawcett (2012), for which exist in the **R** software some functions for its implementation.

The non-stationarity will be here approached simply on a parametric level, through some modifications in the extremes models already used, and not in a semi-parametric context, where this already complex problem is exacerbated.

The data set at hand is the sample of 795 female Static Apena freedivers' best personal records over 180 seconds, set between the years of 2002 and 2014. As such, the way to take on the data is here similar to the one on section 4.1.1.3: use the common data division in yearly blocks, being each new block formed by all the personal competitive best records set in that year. However, as before, this entails that the blocks don't all have the same dimensionality and furthermore, there is no temporal ordering of the subsamples within each block. As such, we are forced to consider the year as our time measure, leaving us with only 13 available blocks.

Figure 4.56 shows the box-plots of our data for each year, and is a parallel representation to the observations plot in Figure 4.33 that allows a better visualization of the behaviour of the samples through time (keeping in mind they have different sizes). It could be argued that a slight upward trend is observable through the years, but this could be simply due to the increase in the subsamples size.

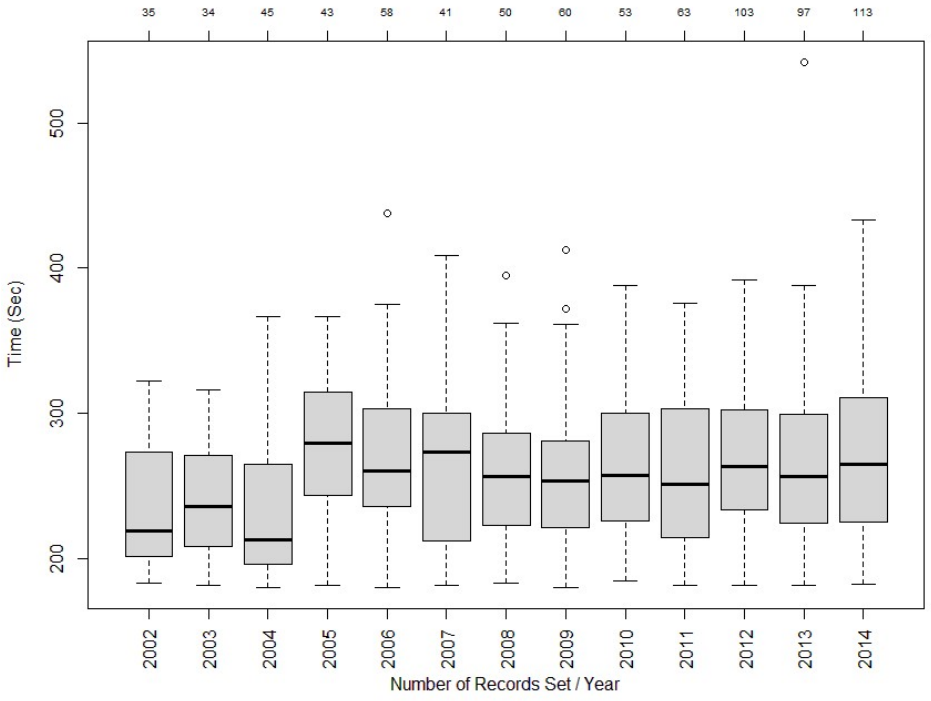

Figure 4.56: Box-plots of the competitive female SA freedivers' individual best records by year

With this division, the non-stationary Block Maxima (or Gumbel) approach – NSBM – is in this case impaired, just as seen in the stationary LO approach for $k = 1$ in section 4.1.1.3, for we are forced to fit the extremes model (stationary or not) to a scarce 13 observations sample. Since time is now a factor to be considered, we cannot proceed as in the stationary Gumbel approach, for which we though of each freediver as its own block, regardless of the year it had set its record in.

Thus, our take on the non-stationary approach to the data consists in testing for the influence of time when fitting the GEV extremal process to the Largest Yearly Observations sample (with the already considered $k = 1, 5, 10, 20$ top observations) to which we will refer as non-stationary LO approach – NSLO –, and then for the fitting of the GP model to the sample of excesses over the chosen threshold $u = 240$, through a non-stationary POT methodology – NSPOT.

Estimation under a significant temporal influence, that is, when the stationarity assumption is thrown, depends on the time variable t, thus the prediction must be made for specific years. As mentioned in the theoretical detailing of this methodology, it is very dangerous to perform long term predictions based on the fit for such small number of years (from 2002 to 2014). Thusly, we limited the estimation performed in this section to a three year period from the sample time span, obtaining predictions of the extreme value parameters of interest for 2015, 2016 and 2017.

4.2.1 Largest Yearly Observations Method

Keeping in mind the analysis performed in section 4.1.1.3 and its results, let us evaluate if there is a significant influence of the factor time in the largest observations registered through the years of the female freedivers best individual records.

In Appendix A.43 can be found the code used for the modeling presented ahead, which is based on the log-likelihood expressions for the Multidimensional GEV model in equations (3.38) and (3.39) but when the core parameters $(\xi(t), \mu(t), \sigma(t))$ can be time dependent. Since our aim is to compare the goodness of the fit with and without the time influence, we have to use here the Multidimensional GEV model for the same number of top yearly observations as has been done before – $k = 1, 5, 10, 20$. As mentioned before, the analysis for $k = 1$, which is equivalent to a NSBM analysis of a 13 observations sample, cannot be taken into much consideration, since there are not enough observations for the fit to be appropriate.

From the observation of Figures 4.33 and 4.56 we see it appears to be some form of an upward trend, that is, the extremal observations seem to be getting more extreme (higher) from the earlier to the most recent years. Thus, we start by considering the possible presence of a linear trend in the location parameter: $\mu(t) = \beta_0 + \beta_1 t$, with t the time integer variable, which as been indexed such that $t = 1$ represents the year of 2002 (following the recommendation of Fawcett (2012)). The resulting estimates for this fit are comprised in Table 4.19, were it is also presented the deviance statistic $\mathbf{L}$ computed between this model with a linear trend in $\mu(t)$ and the completely stationary model (Table 4.15) for each considered top observations set k.

As we know, the value of the deviance statistic allows us to decide if the model with more parameters (in this case, the model with a linear trend in $\mu(t)$) is statistically better than the most parsimonious model, for the significance level α we set here as 5%. As we see in Table 4.19, for every considered k, the corresponding deviance between the stationary and non-stationary model is larger than $\chi^2_{1,1-0.05} = 3.841459$, the χ^2_1 distribution's 95%-quantile. As such, by the decision rule detailed in section 3.1.4, we reject the hypothesis that each pair of models is statistically equivalent, thus concluding that the time cofactor has a statistically significant influence on our

data, which destroys our stationarity assumption. The most suitable model here, for each k, is the non-stationary multidimensional GEV model whose parameters are in Table 4.19.

Table 4.19: Estimates for the Multivariate GEV fit with trend in location to the largest k yearly competitive female SA freedivers' individual best records.

k	*logLikelihood* (*Deviance*)	$\hat{\xi}$ (*se*)	$\hat{\beta}_0$ (*se*)	$\hat{\beta}_1$ (*se*)	$\hat{\sigma}$ (*se*)
1	-63.79594 (11.53305)	0.2885489 (0.3356153)	321.1043746 (11.8179436)	7.7438214 (1.3154429)	23.7220284 (6.8393193)
5	-234.5812 (32.91175)	-0.01680631 (0.08845098)	330.93430368 (9.72136755)	7.12274266 (1.05497291)	28.52258923 (3.83826537)
10	-405.4549 (52.55619)	-0.06601683 (0.06442278)	332.32221015 (8.64455014)	7.21001425 (0.81158755)	29.54966844 (3.42000258)
20	-712.5848 (68.818)	-0.1828081 (0.0319849)	336.9201739 (8.5353593)	7.8279222 (0.8357818)	30.0258537 (2.0668841)

Figure 4.57 is a repetition of the representation in Figure 4.34, but to which were added the trend lines obtained from the fitting of each corresponding model. It is then more clear that there is in fact a significant trend underlying our data, translated here in the location parameter of the fitted models.

A similar procedure that we will omit here but can be found in the mentioned Appendix A.43, considering the set of parameters $(\xi, \mu, \sigma(t))$ with $\sigma(t) = \delta_0 + \delta_1 t$ allowed us to conclude that for each yearly top fraction k fitted, the model with the linear trend in the scale parameter was significantly better than the completely stationary model. An analogous conclusion was drawn regarding the non-stationary models with linear trend in the shape parameter $\xi(t)$ and constant location and scale (μ, σ). It is then clear that a stationary model is not appropriate for fitting our data. But now we are presented with three different types of models, with different types of temporal influence, that we must assess.

Seeing from Figure 4.57 that the trend translated in the location parameter seems quite appropriate to the underlying data, we will built up from this model. That is, consider the Multidimensional GEV model with parameters $(\xi, \mu(t), \sigma(t))$, with the corresponding $\mu(t) = \beta_0 + \beta_1 t$ and $\sigma(t) = \delta_0 + \delta_1 t$, and lets evaluate if this model is statistically significant when compared with the model where only $\mu(t)$ depends on time – for what were computed the new deviance statistics in Table 4.20 ahead. In the referred table can also be seen the full set of estimates for the "larger" model parameters, for each value of k.

In this case, for every k, the deviance statistic is inferior to 3.841459, the χ^2_1 distribution's 95%-quantile, meaning that the hypothesis of equivalence between the two models at hand cannot be rejected. As such, we choose the most parsimonious model where only the location parameter depends on time. Analogous comparisons (here omitted but that can be reproduced by the code in the referred Appendix) were performed between the chosen model and the alternative "larger" model which considered linear trends simultaneously on the location and shape parameters, and the conclusion was always the same: the most appropriate model corresponds to the set

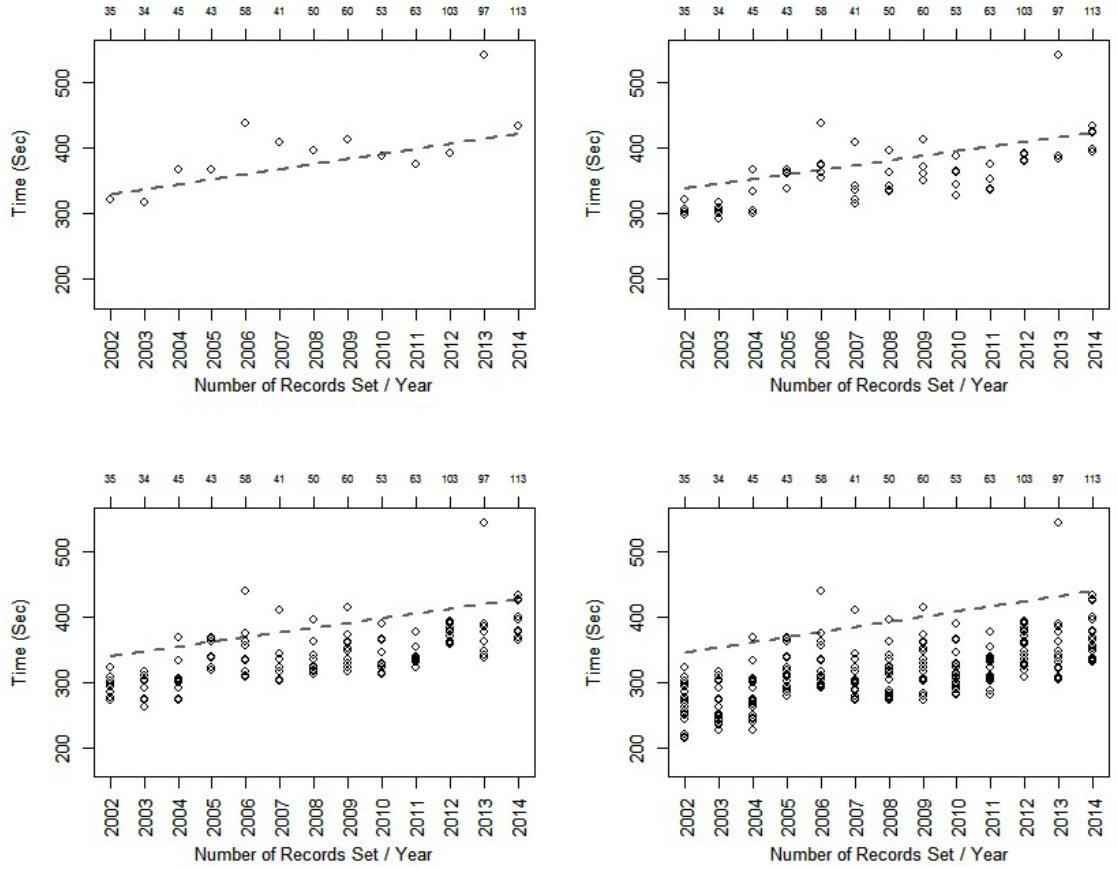

Figure 4.57: Largest 1 (top-left), 5 (to-right), 10 (bottom-left) and 20 (bottom-right) Competitive female SA freedivers' individual best records by year, with fitted trend

of parameters $(\xi, \mu(t), \sigma)$ with the estimates comprised in Table 4.19. This is the model we use for the estimation of the other extreme value indicators of interest for the three year horizon stipulated. The predictions resulting of said estimation can be found in Table 4.21 – see Appendix A.44.

Table 4.20: Estimates for the Multivariate GEV fit with trend in location and scale to the largest k yearly competitive female SA freedivers' individual best records.

k	$logLikelihood$ ($Deviance$)	$\hat{\xi}$ (se)	$\hat{\beta}_0$ (se)	$\hat{\beta}_1$ (se)	$\hat{\delta}_0$ (se)	$\hat{\delta}_1$ (se)
1	-63.42227 (0.7473521)	0.08555764 (0.4318632)	309.89311882 (12.6595087)	9.96292275 (2.8832203)	8.42909563 (14.7111683)	2.77064711 (3.3673206)
5	-233.7558 (1.650966)	-0.03234007 (0.101380)	316.28564516 (13.123730)	9.19771517 (1.969920)	8.61102094 (7.304017)	1.34487340 (1.057284)
10	-405.2116 (0.4866357)	-0.07932348 (0.0687642)	324.45539546 (13.4011823)	8.30853198 (1.7620377)	25.91642944 (5.7594721)	0.44178862 (0.6220557)
20	-712.5846 (0.0002547296)	-0.18417766 (0.03781592)	336.43928735 (14.77895160)	7.88623790 (1.80388376)	29.82697488 (4.67648584)	0.01721869 (0.42522980)

Table 4.21: Three year predictions under the Multivariate GEV fit with trend in location to the largest k yearly competitive female SA freedivers' individual best records.

$Year$	k	$\widehat{P[X > 542]}$	$\widehat{\chi_{0.0001}}$	$\widehat{U(100)}$	$\widehat{x^F}$
2015	1	0.04914832	1519.831	657.3345	-
	5	0.01748190	674.0318	556.9171	2127.788
	10	0.01465513	637.1837	550.4947	880.8705
	20	0.008501086	580.2610	539.9184	610.7591
2016	1	0.05635834	1527.575	665.0783	-
	5	0.02276306	681.1546	564.0399	2134.911
	10	0.02010434	644.3937	557.7047	888.0805
	20	0.015280260	588.0890	547.7463	618.5870
2017	1	0.06496239	1535.318	672.8221	-
	5	0.02958074	688.2773	571.1626	2142.034
	10	0.02737313	651.6037	564.9147	895.2906
	20	0.025881105	595.9169	555.5743	626.4149

Analysing these predictions, there are several aspects that come to our attention. The fit for $k = 1$ (equivalent to the Annual Maximum approach), estimates the EVI as 0.2885489, a very positive and high value, leading to very high estimates of the indicators of interest. This is contrary to all the stationary estimation performed, and is likely the result of fitting a model to a sample of only 13 observations, which is far from appropriate. Thus, we discard these results.

The fit for $k = 5$ estimates the EVI as -0.01680631, a negative value but very close to 0, leading to the highest estimates of the right endpoint seen so far in this case study (around 2130 seconds, i.e., 35 minutes and 30 seconds, for the years of 2015, 2016 and 2017).

It is also evident the positive temporal trend estimated, since the predictions for the parameters are all increasing in value from 2015 through to 2017 (for every k). However, we find these increases to be slight, specially for the right endpoint predictions that change less than 10 seconds from a year to the next.

Another curious aspect is that the predictions for the non-stationary model fitted to the largest 10 yearly observations are very similar to the ones obtained for the stationary model fitted to the largest 5 yearly observations. An analogous relation exists between the non-stationary fit for $k = 20$ and the stationary fit for $k = 10$. The obtained estimates are, in both cases, not very dissonant from the overall estimation performed under the stationarity assumption.

4.2.2 Peaks Over Threshold Method

We will now try to find the temporal influence on the data through another approach – the NSPOT. We have presented in section 3.1.4 two different types of non-stationary POT models: nonparametric non-stationary POT model and the parametric non-stationary POT model, the latter being the one we will focus on this dissertation. Recall that the nonparametric non-stationary POT model consists in dividing the sample in several periods, with different

behaviours. Since we are dealing with the short time span of only 13 years, with no possible break within each year, dividing these 13 years in several different periods for different model fittings would not yield trustworthy results and, furthermore, there is no rule apparent in the data we could use to decide how to divide our sample.

Thus, as stated, we will apply the parametric non-stationary POT model, where the "parametric" term refers only to the time dependence being included in the functional expression of the model parameters. All the results in this section correspond to the code presented in Appendix A.45.

This approach is to be compared to the analysis made in section 4.1.1.2, and as such we will consider the same threshold level $u = 240$ seconds. The observations beyond this mark were plotted by year in Figure 4.58.

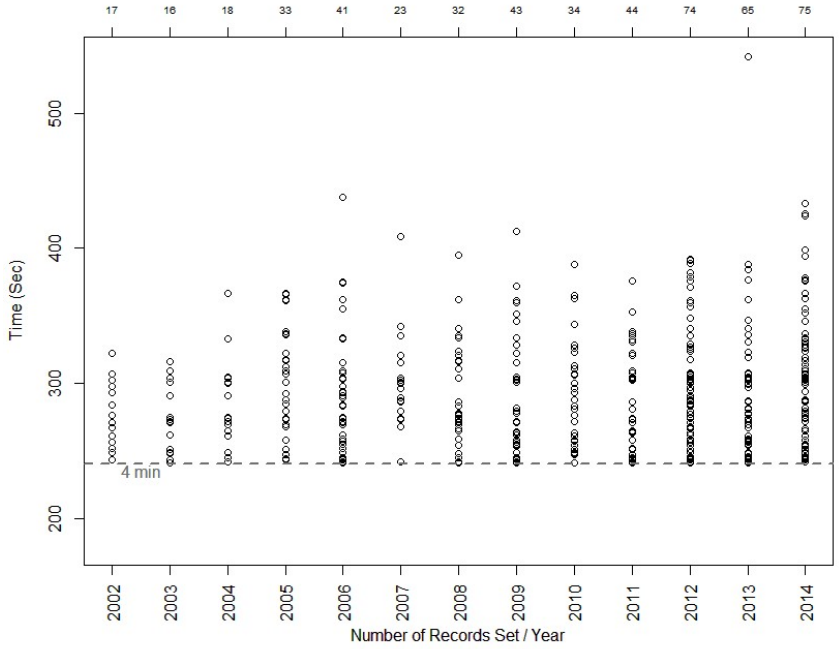

Figure 4.58: Competitive female SA freedivers' individual best records over 240 seconds by year

Recall the stationary GP fit for these data, and consider the ML estimates found in Table 4.14. We will now seek to fit a non-stationary GP model where the dependence of time comes translates only in the scale parameter as $\sigma_u = \exp(\beta_0 + \beta_1 t)$, with t being again the time indicator indexed in $t = 1, \ldots, 13$ corresponding to the years 2002,...,2013. The log-likelihood function for such a model was presented in equation (3.43) in section 3.1.4, and the corresponding parameters estimates are comprised in Table 4.22.

Once again, to perform the LR test it is important to compute the deviance statistic between this model and the completely stationary model, which in this case has the value 9.260549. Being larger than $\chi^2_{1,1-0.05} = 3.841459$, the χ^2_1 distribution's 95%-quantile (since we consider the 5% significance level) we decide on the rejection of the hypothesis that both models are statistically equivalent.

Table 4.22: Estimates for the GP fit with trend in the scale parameter to the competitive female SA freedivers' individual best records.

$logLikelihood\ (Deviance)$	$\hat{\xi}\ (se)$	$\hat{\beta}_0\ (se)$	$\hat{\beta}_1\ (se)$
-2529.987	-0.19037533	3.81956843	0.03273054
(9.260549)	(0.02585366)	(0.09383296)	(0.01015853)

This once more confirms the significative influence the factor time has on our data, and as such we chose to keep the non-stationary model with the estimated parameters in Table 4.22, having an upward linear trend in the log-scale parameter $\sigma_u(t)$. This allows for the prediction of the extreme value indicators of interest for the chosen three years horizon, which results are comprised in Table 4.23.

Table 4.23: Three year predictions under the GP fit with trend in the scale parameter to the competitive female SA freedivers' individual best records.

$Year$	$\widehat{P[X > 542]}$	$\widehat{\chi_{0.0001}}$	$\widehat{U(100)}$	$\widehat{x^F}$
2015	0.0001468666	547.4068	447.4849	618.6278
2016	0.0002750932	557.6348	454.3883	631.2255
2017	0.0004736669	568.2032	461.5215	644.2424

In the predictions for these three years it is clear the estimated upward trend, since the estimates of the indicators increase from year to year. We can compare the predictions for the year 2015 with the stationary estimates, and verify that they are not greatly different, although some disparities can be found, specifically the 2015 predictions being consistently higher than the respective stationary estimates. Even though the right endpoint prediction for 2015 is only 10 seconds higher than its stationary counterpart, the exceedance probability estimate for 2015 is about 10 times greater than said probability estimated under the stationary context. Of course, given the found linear upward trend through time, the 2016 and 2017 predictions for this indicators differ even more from the stationary estimates than those expected for 2015.

The three years taken here for prediction purposes might seem like a very short period of time. However, let us consider the impact that Natalia Molchanova's death, in August 2015, could possibly have in the behaviour of freedivers in future competitions – the trend we estimated could no longer apply. As such, increasing the prediction horizon even further would not be wise.

Chapter 5

Concluding Remarks and Further Topics

In this dissertation our aim was to firstly expose some of the most well known and useful methodologies for analyzing extreme value data, supporting it with a theoretical background, sufficiently detailed for a strong basic understanding of Extreme Value Analysis. This introduction was topped with an illustration of these procedures by applying them to a data set of extreme events in Sports. The sample used in the Case Study refers to competitive best personal records of female freedivers, on the modality Static Apnea. This type of data is naturally extreme and our interest lays with the largest apnea times, thus making sense an EVT approach.

Under the assumptions of the presented models in Chapters 2 and 3, both parametric and semi-parametric methodologies were used on the apnea data. One conclusion they all had in common was pointing to a zero or negative Extreme Value Index, implying we were working under a Gumbel or Weibull max-domain of attraction (although preference was tendentiously given to the Weibull domain). The common sense and physiologically belief that a maximum apnea time humanly possible exists was confirmed by every approach – concluding such a statistical limit exists for competitive female freedivers thusly implies the existence of an equal or probably lower limit for the rest of the female population's apnea capacity. We also verified that the probability the current world record, of the late Molchanova, will be overcome is very slim, around 0.01%.

Specifically considering the stationarity assumption, the Gumbel/Block Maxima, the Peaks over Threshold and the Largest Yearly Observations approaches were used on the competitive apnea data. Although they concur in estimating the EVI in the negative region, the POT approach suggested the tail of the underlying distribution could be even lighter (and shorter) than the BM approach argued for. The LO approach was concordant with both approaches, with conclusions close to those of the BM method when using the top 5 largest yearly observations, and more approximated to the POT conclusions when the top 10 values of each year were considered.

Still ignoring any temporal influence, the semi-parametric was applied on the data and its conclusions regarding the EVI were in line with those of the BM approach. However, it was inconclusive in estimating the right endpoint, though it still indicated the finiteness of this parameter. It was clear that parametric and semi-parametric approaches complement more than

are alternative to each other. This was here evident on the right endpoint estimation, and the semi-parametric approach suggested that the threshold chosen for the parametric POT approach might not have been the most appropriate.

We later learned that assuming the stationarity of the data was abusive, and verified that there was a positive influence of time in the form of a trend that must be considered in the estimation. This implied that the prediction of the parameters was only valid for a short temporal horizon, and resulted on slightly larger estimates than obtained for the stationary methods.

The framework presented in this work allows for an initial take on the most common extremes data, but it is in no way extensive or exclusive, a lot more information on treating extreme data existing in the literature. Further work to be done on the analyzed freediving data includes:

- Considering a *smooth* of the data: since the records were measured in seconds, a lot of repeated values appear on the sample; smoothing the data would ensure more even distribution of the observations throughout the support; this is easily achieved, for example, by adding a pseudo-randomly generated number from a continuous Uniform variable defined in $[-0.5, 0.5]$ seconds to each observation;

- Using a Point Process characterization: it consists in a formulation of the extreme value models based on the point process theory that elegantly combines all the parametric models presented; its likelihood allows for a more natural approach to non-stationarity in threshold models – see Chapter 7 of Coles (2001);

- Applying the Bayesian methodology to the extreme values data: this is a likelihood based methodology that has been gaining terrain in EVT in more recent years, since it allows for the incorporation of external/prior knowledge on the data that could contain relevant information for the analysis; prediction procedures are naturally incorporated on the method and do not depend on the regularity conditions that restrict the ML method – see Bermudez and Turkman (2003) or Chapter 11 of Beirlant et al. (2004);

- Making use of the discarded information in the sample selection: there was a lot of information in the remaining records for each freediver under the personal best, that was ignored in this analysis for the sake of independence in the sample; it would be interesting to find how to incorporate the information of the "repeated readings" for each diver in the analysis and see how it influences the conclusions (since they are lower than the records kept, we suppose this would induce smaller estimated EVI and shorter tailed fitted d.f.);

- Taking a multivariate approach: the complete data provided by AIDA shows many of the divers participate in competitions for several of the freediving modalities; it can be interesting to find if/how the extreme performances in different modalities are related.

We simply lifted the veil of Extreme Value Theory enough to spike the curiosity on the subject and to get some idea of how it can help us comprehend the behaviour of extraordinary events. Its utility was clear in the results of the analysis of the best personal records of SA female competitors and can be attested in a myriad of different types of extreme data not only in Sports, but also in Hydrology, Economics, Finance, Ecology and many more.

Appendix A

R scripts for the Female Freediving Records Case Study

A.1 Data Pre-Selection

```
records_female<-read.csv('Recordes_Anuais_Mulheres.csv', header=T,sep=';')
attach(records_female)
anos<-X.ANO ; recordes<-rp.seg ; nadadores<-as.vector(unique(name))

plot(anos,recordes, type='l',xaxt='n',
     ylab='Records (sec)',xlab='Year', col='grey')
axis(1,at=anos, labels=T, las=3)
for(i in 1:16){
  points(anos[name==nadadores[i]],recordes[name==nadadores[i]])}
text(anos[-c(1,3,5,12,14,15,16)], recordes[-c(1,3,5,12,14,15,16)],
    labels = name[-c(1,3,5,12,14,15,16)], pos = 3, cex=0.7)
text(anos[c(1,3,5)],recordes[c(1,3,5)], labels = name[c(1,3,5)],pos = 4, cex=0.7)
text(anos[c(12,14,15)], recordes[c(12,14,15)], labels = name[c(12,14,15)],
    pos = 1, cex=0.7)
text(anos[16], recordes[16], labels = name[16], pos = 2, cex=0.7)
acf(recordes,main='')
detach(records_female)

female02_14<-read.csv('FEMALE_2002_2014.csv', header=T,sep=';')
attach(female02_14)
anos<-sort(unique(ANO)) ; maxleng<-0
for(i in 1:13){ if(maxleng < length(rpSeg[ANO==2001+i]))
    maxleng<-length(rpSeg[ANO==2001+i])}
numb_obs<-c()
for(i in 1:13){ numb_obs[i]<-length(rpSeg[ANO==2001+i])}
```

```
todos<-matrix(rep(NA,times=13*maxleng),13,maxleng)
for(i in 1:13){ for(j in 1: numb_obs[i])
    todos[i,j]<-sort(rpSeg[ANO==2001+i],decreasing = TRUE)[j]}

plot(anos,todos[,1],ylim=c(0,max(rpSeg)),type='p',xaxt='n',main='',
     ylab='Time (Sec)',xlab='Number of Records Set / Year')
axis(1,at=anos, labels=T, las=3)
axis(3,at=anos, labels=numb_obs, las=1.5, cex.axis=0.6)
for(i in 2:maxleng){points(anos,todos[,i])}
abline(h=120,lty=2)
abline(h=180,lty=6,col='darkgrey')
abline(h=240,lty=3,col='grey50')
text(2002.5,120,'2 min', pos=1)
text(2003.5,180,'3 min', pos=1, col='darkgrey')
text(2004.5,240,'4 min', pos=1, col='grey50' )
detach(female02_14)
```

A.2 Histogram and Box-plot of the Maxima Sample

```
female<-read.csv('FEMALE_2002_2014_sup180.csv', header=T,sep=';')
attach(female)
m<-length(name)
par(mfrow=c(1,2))
hist(rpSeg, xlab='Time (Sec)', main='', col='grey90')
boxplot(rpSeg, horizontal = TRUE, xlab='Time (Sec)', col='grey90')
par(mfrow=c(1,1))
```

A.3 Exponential QQ-plot and sample ME-plot

```
#PLOTING POSITIONS
pp<-(1:m)/(m+1)

#EXPONENCIAL QQPLOT
qq<--log(1-pp)
plot(qq,sort(rpSeg),ylab=expression(y[i:n]),xlab=bquote(-log(1-p[i])),main='')
fitexp<-lm(sort(rpSeg) ~ qq)
summary(fitexp)
cor(sort(rpSeg),-log(1-pp))
abline(fitexp, lty=2,col='grey50', cex=3)
text(5,540,'slope = 52.8396', col='grey50')
text(5,525,'intercept = 208.6084 ', col='grey50')
```

```
text(5.5,180,'correlation = 0.956803 ', col='grey20')

#ME-PLOT
e_kn<-c() ; e_knt<-c()
for(k in 1:(m-1)){ e_kn[k]<-sum(sort(rpSeg)[(m-k+1):m])/k - sort(rpSeg)[m-k]}
for(k in 1:(m-1)){ e_knt[k]<-e_kn[m-k] }
plot(sort(rpSeg)[-c(m, m-1, m-2)],e_knt[-c(m-1,m-2)],type='o',
    xlab=bquote(Threshold (y[m-k:m])), ylab="Mean Excess")
```

A.4 Gumbel QQ-plot

```
qGumbel<-function(x,mu,sigma){mu-sigma*log(-log(x))}
dGumbel<-function(x,mu,sigma){exp(-((x-mu)/sigma+exp(-(x-mu)/sigma)))/sigma}
pGumbel<-function(x,mu,sigma){ exp(-exp((-x+mu)/sigma))}

Qsg<-qGumbel(pp,0,1)
plot(Qsg,sort(rpSeg),ylab=expression(y[i:n]), xlab=bquote(-log(-log(p[i]))))
fitgumb<-lm(sort(rpSeg) ~ Qsg)
summary(fitgumb)
cor(sort(rpSeg),Qsg)
abline(fitgumb, lty=2,col='grey50', cex=3)
text(5,540,'slope = 42.4374', col='grey50')
text(5,525,'intercept = 236.8753 ', col='grey50')
text(5.5,180,'correlation = 0.9903622 ', col='grey20')
```

A.5 Correlation for the GEV QQ-plot

```
library(fExtremes)

correl2<-function(xi,p,dados){ return(cor(dados, qgev(p,xi)))}
bestgama2<-optimize(correl2,p=pp,dados=sort(rpSeg),interval=c(-1,1),maximum = T)
sequ2<-seq(-1,0.55,by=0.0001)
corr2<-sapply(sequ2,correl2,p=pp,dados=sort(rpSeg))
plot(sequ2,corr2,type='l',xlab=bquote(xi),ylab=bquote(correlation(xi)))
abline(v=bestgama2$maximum, col='grey50', lty=2)
abline(h=bestgama2$objective, col='grey50', lty=2)
text(-0.21,0.735,-0.09169215, col='grey50')
text(-0.95,0.99,0.9943272, col='grey50')
```

A.6 GEV QQ-plot

```
Qsgev_pi<-qgev(pp,bestgama2$maximum)
plot(Qsgev_pi,sort(rpSeg),xlab=bquote(((-log(-p[i]))^-hat(xi)-1)/hat(xi)),
    ylab=expression(y[i:n]),main='')
fitgev<-lm(sort(rpSeg) ~ Qsgev_pi)
summary(fitgev)
cor(sort(rpSeg),Qsgev_pi)
abline(fitgev, lty=2,col='grey50', cex=3)
text(4,500,'slope = 47.122', col='grey50')
text(4,485,'intercept = 238.054 ', col='grey50')
text(4,180,'correlation = 0.9943272 ', col='grey20')
text(4,200,bquote(hat(xi) == -0.09169215), col='grey20')
```

A.7 Histogram and Fitted Density Functions for the Gumbel and GEV distributions

```
hist(rpSeg, xlab='Time (Sec)', main='', col='grey90', freq = F, ylim=c(0, 0.0085))
curve(fExtremes::dgev(x,bestgama2$maximum,fitgev$coefficients[1],
    fitgev$coefficients[2]), add = T, lwd=2, col='grey50')
curve(dGumbel(x,fitgumb$coefficients[1],fitgumb$coefficients[2]), add = T, lwd=2)
legend(340,0.007,c("Gumbel (236.8753, 42.4374)","GEV (-0.09169215, 238.054, 47.122)"),
    col=c("black","grey50"),lty = c(1,1),bty='n')
```

A.8 Statistical Choice of Max-Domain of Attraction - BM

```
#ML Gumbel and GEV estimates
library(fitdistrplus)
gumbel_ML_fitd<-fitdist(rpSeg,"gumbel",start=list(loc=236.8753,scale=42.4374))
gev_ML_fitd<-fitdist(rpSeg,"gev", start=list(loc=238.05383,
    scale=47.12241, shape=-0.09169215))

#Hypothesis Tests
Y<-sort(rpSeg)
#H0: xi=0  vs.  H1: xi!=0
  #LRT
l0<- gumbel_ML_fitd$loglik ; l1<- gev_ML_fitd$loglik
D<- -2*(l0-l1)
DD<- D/(1+(2.8/m))  #4.266774
QChi5 <- qchisq(0.95,1)  #3.841459
pval_LRT <- 1-pchisq(DD,1)  #0.03886465
```

```
  #Rao's Score
Z <- (Y-gumbel_ML_fitd$estimate[1])/gumbel_ML_fitd$estimate[2]
Vm <- sum((Z^2)/2 - Z - (Z^2)*exp(-Z)/2 )
VVm2 <- Vm^2 / (2.09797*m)    #2.388051
QChi5 <- qchisq(0.95,1)       #3.841459
pval_SRbi<- 1-pchisq(VVm2,1)  #0.122266
  #LAN
V1<-sum(-Z + (Z^2)*(1-exp(-Z))/2 )
V2<-sum(-1/gumbel_ML_fitd$estimate[2] + (1-exp(-Z))*Z/gumbel_ML_fitd$estimate[2])
V3<-sum(1/gumbel_ML_fitd$estimate[2] - exp(-Z)/gumbel_ML_fitd$estimate[2])
Tm <- (1/3.451)*(1/sqrt(m))*(1.6449*V1 - gumbel_ML_fitd$estimate[2]*0.5066*V2 -
     0.8916*gumbel_ML_fitd$estimate[2]*V3)
TTm <- Tm / 0.6904  #-1.546425
Qnorm975 <- qnorm(0.975)  #1.96
pval_LANbi <- 2-2*pnorm(abs(TTm))    #0.122002
    #Gumbel Statistics - as in Gomes and Fraga Alves (1996)
Gr<-c() ; Gr_ast<-c()
for(r in 1:m){
  rm<-floor((r+1)/2)
  Gr[r]<-(Y[m]-Y[m-rm+1])/(Y[m-rm+1]-Y[m-r+1])
  Gr_ast[r]<-Gr[r]*log(2)-log(rm-1)}
plot(Gr_ast,type='l', xlab="Top Order Statistics", ylab=expression(G[r]^paste('*')),
     mgp=c(2.5,1,0))
crit_points<-read.csv('Pontos Críticos da Estatística de Gumbel Normalizada.csv',
    header=T,sep=';')
attach(crit_points)
alfa25<-X0.025 ; alfa5<-X0.05 ; alfa975<-X0.975
abline(h=alfa25[20],lty=2,col='grey50')
abline(h=alfa975[20],lty=2,col='grey50')
text(790,-1,expression(g[0.025]), col='grey50')
text(790,4,expression(g[0.975]), col='grey50')
rs<-as.numeric(as.vector(crit_points$r[-20]))
points(rs,alfa25[-20], col='grey50', type='l', lwd=1.5)
points(rs,alfa25[-20], col='grey50', pch=19)
points(rs,alfa975[-20],col='grey50', type='l',lwd=1.5)
points(rs,alfa975[-20],col='grey50', pch=19)
legend(150,13,expression(paste("Critical Points of ")*G[r]^paste('*')*
    paste(" at the levels 0.025 and 0.975")),pch=19,col='grey50')

#H0: xi=0  vs.  H1: xi<0
  #Gumbel Statistics - as in Tiago de Oliveira et al. (1984)
gsm <- (Y[m]-Y[ceiling(m/2)])/(Y[ceiling(m/2)]-Y[1])
```

```
betam <- (log(m) + log(log(2))) / (log(log(m)) - log(log(2)))
alpham <- 1/log(log(m))
GSm <- (gsm - betam)/alpham   #1.975694
QGumbel5 <- qgumbel(0.05)     #-1.097189
pval_gumb <- pgumbel(GSm)   #0.8705196
  #Rao's Score
Z <- (Y-gumbel_ML_fitd$estimate[1])/gumbel_ML_fitd$estimate[2]
Vm <- sum((Z^2)/2 - Z - (Z^2)*exp(-Z)/2 )
VVm <- Vm / sqrt(2.09797*m)    #-1.545332
QNorm5 <- qnorm(0.05)         #-1.644854
pval_SRuni<- pnorm(VVm)  #0.061133
  #Teste LAN
V1<-sum(-Z + (Z^2)*(1-exp(-Z))/2 )
V2<-sum(-1/gumbel_ML_fitd$estimate[2] + (1-exp(-Z))*Z/gumbel_ML_fitd$estimate[2])
V3<-sum(1/gumbel_ML_fitd$estimate[2] - exp(-Z)/gumbel_ML_fitd$estimate[2])
Tm <- (1/3.451)*(1/sqrt(m))*(1.6449*V1 - gumbel_ML_fitd$estimate[2]*0.5066*V2 -
     0.8916*gumbel_ML_fitd$estimate[2]*V3)
TTm <- Tm / 0.6904  #-1.546425
QNorm5 <- qnorm(0.05)     #-1.644854
pval_LANuni <- pnorm(TTm)     #0.06100099
    #Gumbel Statistics - as in Gomes and Fraga Alves (1996)
plot(Gr_ast,type='l', xlab="Top Order Statistics", ylab=expression(G[r]^paste('*')),
     mgp=c(2.5,1,0))
abline(h=alfa5[20], lty=2, col='grey50')
text(790,-0.8,expression(g[0.05]), col='grey50')
points(rs,alfa5[-20], col='grey50', type='l', lwd=1.5)
points(rs,alfa5[-20], col='grey50', pch=19)
legend(150,13,expression(paste("Critical Points of ")*G[r]^paste('*')*
    paste(" at the level 0.05")), pch=19,col='grey50')

#Goodness of Fit Tests - H0: F(x)=exp(-exp(-(x-mu_gumb)/sigma_gumb))
goodness_gumbel<-gofstat(gumbel_ML_fitd)
Dm<-sqrt(m)*0.04779257      #1.347547
Wm2<-0.45724238*(1+0.2/sqrt(m))   #0.4604857
Am2<-3.62113458*(1+0.2/sqrt(m))   #3.64682
```

A.9 Gumbel Fitting to the Sample of Maxima

```
library(ismev) ; library(evd) ; max(rpSeg) #542

#Preliminary
  gum_prob542_pre<-1-pGumbel(542,236.8753,42.4374)
```

```
  gum_quant0.0001_pre<-qGumbel(0.9999,236.8753,42.4374)
  gum_quant0.01_pre<-qGumbel(0.99,236.8753,42.4374)

#ML
    #fitdistrplus
  gumbel_ML_fitd<-fitdist(rpSeg,"gumbel", start=list(loc=236.8753,scale=42.4374))
  plot(gumbel_ML_fitd)
  gum_prob542_ML_fitd<-1-pgumbel(542,gumbel_ML_fitd$estimate[1],
        gumbel_ML_fitd$estimate[2])
  gum_quant0.0001_ML_fitd<-qGumbel(0.9999,gumbel_ML_fitd$estimate[1],
        gumbel_ML_fitd$estimate[2])
  gum_quant0.01_ML_fitd<-qGumbel(0.99,gumbel_ML_fitd$estimate[1],
        gumbel_ML_fitd$estimate[2])
    #fExtremes
  gumbel_ML_fE<-gumbelFit(rpSeg,type="mle")
  summary(gumbel_ML_fE)
  gum_prob542_ML_fE<-1-pGumbel(542,235.69408,44.87459)
  gum_quant0.0001_ML_fE<-qGumbel(0.9999,235.69408,44.87459)
  gum_quant0.01_ML_fE<-qGumbel(0.99,235.69408,44.87459)
    #ismev
  gumbel_ML_ism<-gum.fit(rpSeg)
  gum.diag(gumbel_ML_ism)
  gum_prob542_ML_ism<-1-pGumbel(542,gumbel_ML_ism$mle[1],gumbel_ML_ism$mle[2])
  gum_quant0.0001_ML_ism<-qGumbel(0.9999,gumbel_ML_ism$mle[1],gumbel_ML_ism$mle[2])
  gum_quant0.01_ML_ism<-qGumbel(0.99,gumbel_ML_ism$mle[1],gumbel_ML_ism$mle[2])

    #evd
  gumbel_ML_evd<-fgev(rpSeg, shape=0)
  par(mfrow=c(2,2))
  plot(gumbel_ML_evd)
  par(mfrow=c(1,1))
  gum_prob542_ML_evd<-1-pGumbel(542,gumbel_ML_evd$estimate[1],
        gumbel_ML_evd$estimate[2])
  gum_quant0.0001_ML_evd<-qGumbel(0.9999,gumbel_ML_evd$estimate[1],
        gumbel_ML_evd$estimate[2])
  gum_quant0.01_ML_evd<-qGumbel(0.99,gumbel_ML_evd$estimate[1],
        gumbel_ML_evd$estimate[2])

#PWM
M100=mean(rpSeg) ; yy<-c()
for(i in 1:m) { yy[i]=(i-1)/(m-1)*sort(rpSeg)[i]}
M110=mean(yy)
```

```
beta_MPP=(2*M110-M100)/log(2)
euler=0.57721
mu_MPP=M100-euler*beta_MPP
cat("mu:",mu_MPP,"  beta:",beta_MPP)
  gum_prob542_PWM_mao<-1-pGumbel(542,mu_MPP,beta_MPP)
  gum_quant0.0001_PWM_mao<-qGumbel(0.9999,mu_MPP,beta_MPP)
  gum_quant0.01_PWM_mao<-qGumbel(0.99,mu_MPP,beta_MPP)

#Confidence Intervals
    #parameters
  Gumb_ML_CI<-confint(profile(gumbel_ML_evd, conf = 0.95))
  Gumb_ML_CI_loc<-c(Gumb_ML_CI[1],Gumb_ML_CI[3])
  Gumb_ML_CI_scale<-c(Gumb_ML_CI[2],Gumb_ML_CI[4])
  par(mfrow=c(1,2))
  plot(profile(gumbel_ML_evd,which='loc', conf=0.95))
  abline(v=Gumb_ML_CI_loc[1],lty=2,col='grey50')
  abline(v=Gumb_ML_CI_loc[2],lty=2,col='grey50')
  abline(v=gumbel_ML_evd$estimate[1],lty=2,col='grey80')
  text(233,-4273.8,round(Gumb_ML_CI_loc[1],4),col='grey50')
  text(238.4,-4273.8,round(Gumb_ML_CI_loc[2],4),col='grey50')
  text(gumbel_ML_evd$estimate[1],-4273.8,round(gumbel_ML_evd$estimate[1],4),
       col='grey80')
  plot(profile(gumbel_ML_evd,which='scale', conf=0.95))
  abline(v=Gumb_ML_CI_scale[1],lty=2,col='grey50')
  abline(v=Gumb_ML_CI_scale[2],lty=2,col='grey50')
  abline(v=gumbel_ML_evd$estimate[2],lty=2,col='grey80')
  text(42.9,-4273.95,round(Gumb_ML_CI_scale[1],4),col='grey50')
  text(47,-4273.95,round(Gumb_ML_CI_scale[2],4),col='grey50')
  text(gumbel_ML_evd$estimate[2],-4273.95,round(gumbel_ML_evd$estimate[2],4),
       col='grey80')
  par(mfrow=c(1,1))

    #q(0.0001) and U(100)
  log_L<-c();rl_fix<-c()
  for(i in 1:60000){
    rli<-620+i/1000
    logL<-function(sigma,y){
      lamb<-rli+sigma*log(-log(0.9999))
      logl<--m*log(sigma)-sum(exp(-(y-lamb)/sigma))-sum((y-lamb)/sigma)
      return(logl)}
    out<-optimize(logL,interval=c(30,60),y=rpSeg, maximum=T)
    rl_fix[i]<-rli
```

```
    log_L[i]<-out$objective}
  plot(rl_fix,log_L,type='l', xlab="Quantile 0.9999",ylab="profile log-likelihood")
  q=qchisq(0.95,df=1)
  abline(h=max(log_L))
  abline(h=max(log_L)-q/2)
  IC1<-rl_fix[round(log_L,4)==round(max(log_L)-q/2,3)]
  IC2<-rl_fix[round(log_L,4)==round(max(log_L)-q/2,4)]
  abline(v=IC1,lty=2,col='grey50')
  abline(v=IC2,lty=2,col='grey50')
  abline(v=rl_fix[log_L==max(log_L)],lty=2,col='grey80')
  text(630,-4274,IC1,col='grey50')
  text(670,-4274,IC2,col='grey50')
  text(rl_fix[log_L==max(log_L)],-4274,rl_fix[log_L==max(log_L)],col='grey80')
  log_L<-c();rl_fix<-c()
  for(i in 1:35000){
    rli<-425+i/1000
    logL<-function(sigma,y){
      lamb<-rli+sigma*log(-log(0.99))
      logl<--m*log(sigma)-sum(exp(-(y-lamb)/sigma))-sum((y-lamb)/sigma)
      return(logl)}
    out<-optimize(logL,interval=c(30,60),y=rpSeg, maximum=T)
    rl_fix[i]<-rli
    log_L[i]<-out$objective}
  plot(rl_fix,log_L,type='l', xlab="U(100)",ylab="profile log-likelihood")
  q=qchisq(0.95,df=1)
  abline(h=max(log_L))
  abline(h=max(log_L)-q/2)
  IC1<-rl_fix[round(log_L,4)==round(max(log_L)-q/2,3)]
  abline(v=IC1,lty=2,col='grey50')
  abline(v=rl_fix[log_L==max(log_L)],lty=2,col='grey80')
  text(432,-4274,IC1[1],col='grey50')
  text(453,-4274,IC1[2],col='grey50')
  text(rl_fix[log_L==max(log_L)],-4274,rl_fix[log_L==max(log_L)],col='grey80')
```

A.10 GEV Fitting to the Sample of Maxima

```
#Preliminar
  gev_prob542_pre<-1-pgev(542,238.05383,47.12241,-0.09169215)
  gev_quant0.0001_pre<-qgev(0.9999,238.05383,47.12241,-0.09169215)
  gev_quant0.01_pre<-qgev(0.99,238.05383,47.12241,-0.09169215)
  endpoint_pre<-238.05383-47.12241/(-0.09169215)
```

```
#ML
    #fitdistrplus
  gev_ML_fitd<-fitdist(rpSeg,"gev", start=list(loc=238.05383,scale=47.12241,
        shape=-0.09169215))
  plot(gev_ML_fitd)
  gev_prob542_ML_fitd<-1-pgev(542,gev_ML_fitd$estimate[1],
        gev_ML_fitd$estimate[2],gev_ML_fitd$estimate[3])
  gev_quant0.0001_ML_fitd<-qgev(0.9999,gev_ML_fitd$estimate[1],
        gev_ML_fitd$estimate[2],gev_ML_fitd$estimate[3])
  gev_quant0.01_ML_fitd<-qgev(0.99,gev_ML_fitd$estimate[1],
        gev_ML_fitd$estimate[2],gev_ML_fitd$estimate[3])
  endpoint_ML_fitd<-gev_ML_fitd$estimate[1]-
        gev_ML_fitd$estimate[2]/gev_ML_fitd$estimate[3]
    #fExtremes
  gev_ML_fE<-gevFit(rpSeg,type="mle")
  summary(gev_ML_fE)
  gev_prob542_ML_fE<-1-pgev(542,237.20621763,45.87758286,-0.06275239)
  gev_quant0.0001_ML_fE<-qgev(0.9999,237.20621763,45.87758286,-0.06275239)
  gev_quant0.01_ML_fE<-qgev(0.99,237.20621763,45.87758286,-0.06275239)
  endpoint_ML_fE<-237.20621763-45.87758286/(-0.06275239)
    #ismev
  gev_ML_ism<-gev.fit(rpSeg)
  gev.diag(gev_ML_ism)
  gev_prob542_ML_ism<-1-pgev(542,gev_ML_ism$mle[1],gev_ML_ism$mle[2],
        gev_ML_ism$mle[3])
  gev_quant0.0001_ML_ism<-qgev(0.9999,gev_ML_ism$mle[1],gev_ML_ism$mle[2],
        gev_ML_ism$mle[3])
  gev_quant0.01_ML_ism<-qgev(0.99,gev_ML_ism$mle[1],gev_ML_ism$mle[2],
        gev_ML_ism$mle[3])
  endpoint_ML_ism<-gev_ML_ism$mle[1]-gev_ML_ism$mle[2]/gev_ML_ism$mle[3]
    #evd
  gev_ML_evd<-fgev(rpSeg)
  par(mfrow=c(2,2)) ; plot(gev_ML_evd)
  par(mfrow=c(1,1))
  gev_prob542_ML_evd<-1-pgev(542,gev_ML_evd$estimate[1],
        gev_ML_evd$estimate[2],gev_ML_evd$estimate[3])
  gev_quant0.0001_ML_evd<-qgev(0.9999,gev_ML_evd$estimate[1],
        gev_ML_evd$estimate[2],gev_ML_evd$estimate[3])
  gev_quant0.01_ML_evd<-qgev(0.99,gev_ML_evd$estimate[1],
        gev_ML_evd$estimate[2],gev_ML_evd$estimate[3])
  endpoint_ML_evd<-gev_ML_evd$estimate[1]-
        gev_ML_evd$estimate[2]/gev_ML_evd$estimate[3]
```

```
#PWM
M100=mean(rpSeg); yy<-c(); yyy<-c()
for(i in 1:m) {
  yy[i]=(i-1)/(m-1)*sort(rpSeg)[i]
  yyy[i]=(i-1)*(i-2)/((m-1)*(m-2))*sort(rpSeg)[i]}
M110=mean(yy) ; M120=mean(yyy)
h<-function(g){(3*M120-M100)/(2*M110-M100)-(3^g-1)/(2^g-1)}
g<-uniroot(h,lower=-0.3,upper=0.3)$root
beta_MPP=g*(2*M110-M100)/(gamma(1-g)*(2^g-1))
mu_MPP=M100+beta_MPP*(1-gamma(1-g))/g
cat("xi:",g,"  mu:",mu_MPP,"  beta:",beta_MPP)
  gev_prob542_PWM_mao<-1-pgev(542,mu_MPP,beta_MPP,g)
  gev_quant0.0001_PWM_mao<-qgev(0.9999,mu_MPP,beta_MPP,g)
  gev_quant0.01_PWM_mao<-qgev(0.99,mu_MPP,beta_MPP,g)
  endpoint_PWN_mao<-mu_MPP-beta_MPP/g

#Confidence Intervals
    #parameters
  GEV_ML_CI<-confint(profile(gev_ML_evd, conf = 0.95))
  GEV_ML_CI_loc<-c(GEV_ML_CI[1],GEV_ML_CI[4])
  GEV_ML_CI_scale<-c(GEV_ML_CI[2],GEV_ML_CI[5])
  GEV_ML_CI_shape<-c(GEV_ML_CI[3],GEV_ML_CI[6])
  par(mfrow=c(1,2))
  plot(profile(gev_ML_evd,which='loc', conf=0.95))
  abline(v=GEV_ML_CI_loc[1],lty=2,col='grey50')
  abline(v=GEV_ML_CI_loc[2],lty=2,col='grey50')
  abline(v=gev_ML_evd$estimate[1],lty=2,col='grey80')
  text(234.3,-4271.65,round(GEV_ML_CI_loc[1],4),col='grey50')
  text(240.2,-4271.65,round(GEV_ML_CI_loc[2],4),col='grey50')
  text(gev_ML_evd$estimate[1],-4271.65,round(gev_ML_evd$estimate[1],4),
       col='grey80')
  plot(profile(gev_ML_evd,which='scale', conf=0.95))
  abline(v=GEV_ML_CI_scale[1],lty=2,col='grey50')
  abline(v=GEV_ML_CI_scale[2],lty=2,col='grey50')
  abline(v=gev_ML_evd$estimate[2],lty=2,col='grey80')
  text(43.8,-4271.78,round(GEV_ML_CI_scale[1],4),col='grey50')
  text(48.15,-4271.78,round(GEV_ML_CI_scale[2],4),col='grey50')
  text(gev_ML_evd$estimate[2],-4271.78,round(gev_ML_evd$estimate[2],4),col='grey80')
  plot(profile(gev_ML_evd,which='shape', conf=0.95))
  abline(v=GEV_ML_CI_shape[1],lty=2,col='grey50')
  abline(v=GEV_ML_CI_shape[2],lty=2,col='grey50')
  abline(v=gev_ML_evd$estimate[3],lty=2,col='grey80')
```

```
  text(-0.103,-4271.78,round(GEV_ML_CI_shape[1],4),col='grey50')
  text(-0.013,-4271.78,round(GEV_ML_CI_shape[2],4),col='grey50')
  text(gev_ML_evd$estimate[3],-4271.78,round(gev_ML_evd$estimate[3],4),col='grey80')
  par(mfrow=c(1,1))

    #q(0.0001) and U(100)
  gevrlevelPlot(gev_ML_fE,kBlocks = 10000,ci = 0.95)
  abline(v=508.204,lty=2,col='grey50')
  abline(v=644.6507,lty=2,col='grey50')
  abline(v=558.127,lty=2,col='grey80')
  text(520,-4271.85,508.204,col='grey50')
  text(630,-4271.85,644.6507,col='grey50')
  text(558.127,-4271.85,558.127,col='grey80')
  IC<-gev.prof(gev_ML_ism, m=100, 400,455, nint=1000)
  abline(v=gev_quant0.01_ML_ism,lty=2,col='grey80')
  abline(v=IC,lty=2,col='grey50')
  abline(v=gev_quant0.01_ML_ism,lty=2,col='grey80')
  text(407,-4271.9,IC1[1],col='grey50')
  text(441,-4271.9,IC1[2],col='grey50')
  text(gev_quant0.01_ML_ism,-4271.9,round(gev_quant0.01_ML_ism,4),
        col='grey80')
```

A.11 POT Choice of Threshold

```
library(evir)
m<-length(name)

#ME-plots
  e_kn<-c()
  for(k in 1:(m-1)){
    e_kn[k]<-sum(sort(rpSeg)[(m-k+1):m])/k - sort(rpSeg)[m-k]}
  e_knt<-c()
  for(k in 1:(m-1)){ e_knt[k]<-e_kn[m-k]}
  par(mfrow=c(1,2))
    #SEPARATION AT 5 MINUTES = 300 SECONDS
  plot(sort(rpSeg)[-c(m, m-1, m-2)],e_knt[-c(m-1,m-2)],type='o',
        xlab=bquote(Threshold (y[n-k:n])), ylab="Mean Excess")
  abline(v=300,col='grey20', lty=2)
  max(which(sort(rpSeg)<=300))
  fitline3<-lm(e_knt[1:595]~sort(rpSeg)[1:595])
  abline(fitline3,col='grey20')
  fitline6<-lm(e_knt[596:(m-3)]~sort(rpSeg)[596:(m-3)])
```

```
  abline(fitline6,col='grey60')
  cor3<-cor(e_knt[1:595],sort(rpSeg)[1:595])
  cor6<-cor(e_knt[596:(m-3)],sort(rpSeg)[596:(m-3)])
  text(225,70,round(cor3,4), col='grey20')
  text(410,28,round(cor6,4), col='grey60')
    #SEPARATION AT 4 MINUTES = 240 SECONDS
  plot(sort(rpSeg)[-c(m, m-1, m-2)],e_knt[-c(m-1,m-2)],type='o',
        xlab=bquote(Threshold (y[n-k:n])), ylab="Mean Excess")
  abline(v=240,col='grey20', lty=2)
  a<-max(which(sort(rpSeg)<=240))
  fitline3<-lm(e_knt[1:a]~sort(rpSeg)[1:a])
  abline(fitline3,col='grey20')
  fitline6<-lm(e_knt[(a+1):(m-3)]~sort(rpSeg)[(a+1):(m-3)])
  abline(fitline6,col='grey60')
  cor33<-cor(e_knt[1:a],sort(rpSeg)[1:a])
  cor66<-cor(e_knt[(a+1):(m-3)],sort(rpSeg)[(a+1):(m-3)])
  text(225,70,round(cor33,4), col='grey20')
  text(300,43,round(cor66,4), col='grey60')
  par(mfrow=c(1,1))

#Parameters Estimates against threshold
  gpd.fitrange(rpSeg,200,400,nint=100,show = TRUE)
```

A.12 Plot of Exceedances over $u = 240$ seconds

```
u1<-240 ; exceedances1<-rpSeg[rpSeg>u1] ; Nu1<-length(exceedances1) #515
excesses1<-exceedances1-u1 ; M<-1:m
plot(M,rpSeg,type='h', xlab='Index',ylab='Time (Sec)', col='grey70')
abline(h=u1, lwd=2)
points(M[rpSeg>u1],exceedances1, cex=0.7,bg='black', pch=21)
```

A.13 Exponential QQ-plot

```
#PLOTING POSITIONS
pppot1<-(1:Nu1)/(Nu1+1)

#EXPONENCIAL QQPLOT
qqpot1<--log(1-pppot1)
plot(qqpot1,sort(excesses1),ylab=expression(y[i:N[240]]),
     xlab=bquote(-log(1-p[i])), main='')
fitexppot1<-lm(sort(excesses1)~qqpot1-1)
```

```
summary(fitexppot1)
cor(sort(excesses1),qqpot1)
abline(fitexppot1, lty=2,col='grey50', cex=3)
text(5,270,'slope = 47.2122', col='grey50')
text(5.5,0,'correlation = 0.983852 ', col='grey20')
```

A.14 Correlation for the GP QQ-plot

```
correlat<-function(xi,p,dados){ return(cor(dados, qgpd(p,xi)))}
bestgama_gp<-optimize(correlat,p=pppot1,dados=sort(excesses1), interval=c(-1,1),
     maximum = T)
sequ_gp<-seq(-1,0.55,by=0.0001)
correlations<-sapply(sequ_gp,correlat,p=pppot1,dados=sort(excesses1))
plot(sequ_gp,correlations,type='l',xlab=bquote(xi),ylab=bquote(correlation(xi)))
abline(v=bestgama_gp$maximum, col='grey50', lty=2)
abline(h=bestgama_gp$objective, col='grey50', lty=2)
text(-0.21,0.80,-0.1967652, col='grey50')
text(-0.95,0.989,0.9927699, col='grey50')
```

A.15 GP QQ-plot

```
QQgpd<-qgpd(pppot1,bestgama_gp$maximum)
plot(QQgpd,sort(excesses1),ylab=expression(y[i:N[240]]),
        xlab=bquote(((1-p[i])^-hat(xi)-1)/hat(xi)), main='')
fitgpd<-lm(sort(excesses1) ~ QQgpd-1)
summary(fitgpd)
cor(sort(excesses1), QQgpd)
abline(fitgpd, lty=2,col='grey50', cex=3)
text(3,210,'slope = 60.8800', col='grey50')
text(3,0,'correlation = 0.9927699 ', col='grey20')
text(3,18,bquote(hat(xi) == -0.1967652), col='grey20')
```

A.16 Statistical Choice of Max-Domain of Attraction - POT

```
#ML Exponential and GP estimates
exp_ML_fitdp<-fitdist(excesses1,'exp',start=list(rate=1/47.2122))
  exp_sigma_ML<-1/exp_ML_fitdp$estimate
GPd_ML_fitd<-fitdist(excesses1,"gpd", start=list(xi=-0.1967652, beta=60.8800))
```

```
#Hypothesis Tests
W<-sort(exceedances1)
#H0: xi=0  vs.  H1: xi!=0
  #LRT
l0<- exp_ML_fitdp$loglik ; l1<- GPd_ML_fitd$loglik
L<- -2*(l0-l1)
LL<- L/(1+(4/Nu1))       #18.41124
QChi5 <- qchisq(0.95,1)  #3.841459
pval_LRT <- 1-pchisq(LL,1)  #1.78005e-05
  #T{N_u} Test Statistic - as in Marohn (2000)
Sw2<-(Nu1-1)*var(exceedances1)/Nu1
Wbar<-mean(exceedances1)
Tm <- (1/2)*(Sw2/((Wbar-u1)^2) -1)
TTm <- sqrt(Nu1)*Tm              #-3.702529
Qnorm975 <- qnorm(0.975)         #1.96
pval_LANbi <- 2-2*pnorm(abs(TTm))  #0.0002134613

#H0: xi=0  vs.  H1: xi<0
  #T{N_u} Test Statistic - as in Marohn (2000)
Sw2<-(Nu1-1)*var(exceedances1)/Nu1
Wbar<-mean(exceedances1)
Tm <- (1/2)*(Sw2/((Wbar-u1)^2) -1)
TTm <- sqrt(Nu1)*Tm          #-3.702529
QNorm5 <- qnorm(0.05)        #-1.644854
pval_LANuni <- pnorm(TTm)    #0.0001067306
  #G{N_u} Test Statistic - as in Gomes and van Monfort (1986)
W<-sort(exceedances1)
GNu<-W[Nu1]/W[ceiling(Nu1/2)]
GGNu<-log(2)*GNu-log(Nu1)     #-4.916655
QGumbel5 <- qgumbel(0.05)     #-1.097189
pval_gumb <- pgumbel(GGNu)    #5.002582e-60

#Goodness of Fit Test for the Exponential model
NU1<-1:Nu1
    #Kolmogorov-Smirnov
ks<-max(abs(1-exp(-sort(excesses1)/exp_sigma_ML)-NU1/Nu1),abs(1-exp(-sort(excesses1)/
        exp_sigma_ML)-(NU1-1)/Nu1))     #0.09562227
CV_ks_1<-1.25/sqrt(Nu1)                 #0.05508158

#Goodness of Fit Tests for the GP model
sigmaism<-GPd_ML_ism$mle[1]
xiism<-GPd_ML_ism$mle[2]
```

```
    #Cramér-von Mises
Wm2<-sum((evd::pgpd(sort(excesses1),scale=sigmaism, shape=xiism)-
        (2*NU1-1)/(2*Nu1))^2)+1/(12*Nu1)
    #Anderson-Darling
Am2<--Nu1-(1/Nu1)*sum((2*NU1-1)*log(evd::pgpd(sort(excesses1),
        scale=sigmaism, shape=xiism))+(2*Nu1+1-2*NU1)*log(1-evd::pgpd(sort(excesses1),
        scale=sigmaism, shape=xiism)))
```

A.17 Exponential Fitting to the Excesses Sample

```
#Preliminar
exp_prob542_pre<-(Nu1/m)*exp(-(542-u1)/47.2122)
exp_quant0.0001_pre<-u1-47.2122*log(m*0.0001/Nu1)
exp_quant0.01_pre<-u1-47.2122*log(m*0.01/Nu1)

#ML
    #fitdistrplus
  exp_ML_fitdp<-fitdist(excesses1,'exp',start=list(rate=1/47.2122))
  exp_sigma_ML<-1/exp_ML_fitdp$estimate
  plot(exp_ML_fitdp)
  exp_prob542_ML<-(Nu1/m)*exp(-(542-u1)/exp_sigma_ML)
  exp_quant0.0001_ML<-u1-exp_sigma_ML*log(m*0.0001/Nu1)
  exp_quant0.01_ML<-u1-exp_sigma_ML*log(m*0.01/Nu1)

#Confidence Intervals
    #parameter
  logLike<-function(y,sigma){return(-length(y)*log(sigma)-1/sigma*sum(y))}
  profLik_scale<-optimize(logLike,y=excesses1,lower=45,upper=60,maximum=T)
  sigma <- seq(46,57, .00001)
  plot(sigma,logLike(excesses1,sigma), type = "l",main="Profile log-likelihood
        of Scale", xlab="scale",ylab="profile log-likelihood" )
  q=qchisq(0.95,df=1)
  abline(h=profLik_scale$objective)
  abline(h=profLik_scale$objective-q/2)
  IC<-sigma[round(logLike(excesses1,sigma),5)==round(profLik_scale$objective-q/2,5)]
  abline(v=IC[1],lty=2,col='grey50')
  abline(v=IC[2],lty=2,col='grey50')
  abline(v=profLik_scale$maximum,lty=2,col='grey80')
  text(47.9,-2546,IC[1],col='grey50')
  text(55.5,-2546,IC[2],col='grey50')
  text(profLik_scale$maximum,-2546,round(profLik_scale$maximum,4),col='grey80')
```

```
  #q(0.0001) and U(100)
logLike2<- function(y,quantil) {
  sigma<-(u1-quantil)/log(0.0001*m/Nu1)
  return( -length(y) * log(sigma) - 1/sigma * sum(y))}
profLik_quant<-optimize(logLike2,y=excesses1,lower=400,upper=900,maximum=T)
quanti<-seq(640,750,0.0001)
plot(quanti,logLike2(excesses1,quanti), type = "l",main="Profile log-likelihood
      of Quantile 0.9999", xlab="Quantile 0.9999",ylab="profile log-likelihood" )
q=qchisq(0.95,df=1)
abline(h=profLik_quant$objective)
abline(h=profLik_quant$objective-q/2)
IC<-quanti[round(logLike2(excesses1,quanti),5)==round(profLik_quant$objective-q/2,5)]
abline(v=IC[1],lty=2,col='grey50')
abline(v=IC[2],lty=2,col='grey50')
abline(v=profLik_quant$maximum,lty=2,col='grey80')
text(660,-2546,IC[1],col='grey50')
text(725,-2546,IC[2],col='grey50')
text(profLik_quant$maximum,-2546,round(profLik_quant$maximum,4),col='grey80')
logLike3<- function(y,quantil) {
  sigma<-(u1-quantil)/log(0.01*m/Nu1)
  return( -length(y) * log(sigma) - 1/sigma * sum(y))}
profLik_ret<-optimize(logLike3,y=excesses1,lower=400,upper=900,maximum=T)
ret<-seq(435,485,0.0001)
plot(ret,logLike3(excesses1,ret), type = "l",main="Profile log-likelihood
      of Return Level U(100)", xlab="U(100)",ylab="profile log-likelihood" )
q=qchisq(0.95,df=1)
abline(h=profLik_ret$objective)
abline(h=profLik_ret$objective-q/2)
IC<-ret[round(logLike3(excesses1,ret),5)==round(profLik_ret$objective-q/2,5)]
abline(v=IC[1],lty=2,col='grey50')
abline(v=IC[2],lty=2,col='grey50')
abline(v=profLik_ret$maximum,lty=2,col='grey80')
text(440,-2546,IC[1],col='grey50')
text(470.5,-2546,IC[2],col='grey50')
text(profLik_ret$maximum,-2546,round(profLik_ret$maximum,4),col='grey80')
```

A.18 GP Fitting to the Excesses Sample

```
#Preliminar
  GPd_prob542_pre<-(Nu1/m)*(1+(-0.1967652)*(542-u1)/60.8800)^(-1/(-0.1967652))
  GPd_quant0.0001_pre<-u1+60.8800/(-0.1967652)*((m*0.0001/Nu1)^-(-0.1967652)-1)
  GPd_quant0.01_pre<-u1+60.8800/(-0.1967652)*((m*0.01/Nu1)^-(-0.1967652)-1)
```

```
  GPd_endpoint_pre<-u1-60.8800/(-0.1967652)

#ML
    #fitdistrplus
  GPd_ML_fitd<-fitdist(excesses1,"gpd", start=list(xi=-0.1967652, beta=60.8800))
  plot(GPd_ML_fitd)
  GPd_prob542_ML_fitd<-(Nu1/m)*(1+GPd_ML_fitd$estimate[1]*(542-u1)/
        GPd_ML_fitd$estimate[2])^(-1/GPd_ML_fitd$estimate[1])
  GPd_quant0.0001_ML_fitd<-u1+GPd_ML_fitd$estimate[2]/GPd_ML_fitd$estimate[1]*
        ((m*0.0001/Nu1)^-GPd_ML_fitd$estimate[1]-1)
  GPd_quant0.01_ML_fitd<-u1+GPd_ML_fitd$estimate[2]/GPd_ML_fitd$estimate[1]*
        ((m*0.01/Nu1)^-GPd_ML_fitd$estimate[1]-1)
  GPd_endpoint_ML_fitd<-u1-GPd_ML_fitd$estimate[2]/GPd_ML_fitd$estimate[1]
    #evir
  GPd_ML_evir<-gpd(rpSeg,u1)
  GPd_prob542_ML_evir<-(Nu1/m)*(1+GPd_ML_evir$par.ests[1]*(542-u1)/
        GPd_ML_evir$par.ests[2])^(-1/GPd_ML_evir$par.ests[1])
  GPd_quant0.0001_ML_evir<-u1+GPd_ML_evir$par.ests[2]/GPd_ML_evir$par.ests[1]*
        ((m*0.0001/Nu1)^-GPd_ML_evir$par.ests[1]-1)
  GPd_quant0.01_ML_evir<-u1+GPd_ML_evir$par.ests[2]/GPd_ML_evir$par.ests[1]*
        ((m*0.01/Nu1)^-GPd_ML_evir$par.ests[1]-1)
  GPd_endpoint_ML_evir<-u1-GPd_ML_evir$par.ests[2]/GPd_ML_evir$par.ests[1]
    #ismev
  GPd_ML_ism<-gpd.fit(rpSeg,u1)
  gpd.diag(GPd_ML_ism)
  GPd_prob542_ML_ism<-(Nu1/m)*(1+GPd_ML_ism$mle[2]*(542-u1)/
        GPd_ML_ism$mle[1])^(-1/GPd_ML_ism$mle[2])
  GPd_quant0.0001_ML_ism<-u1+GPd_ML_ism$mle[1]/GPd_ML_ism$mle[2]*
        ((m*0.0001/Nu1)^-GPd_ML_ism$mle[2]-1)
  GPd_quant0.01_ML_ism<-u1+GPd_ML_ism$mle[1]/GPd_ML_ism$mle[2]*
        ((m*0.01/Nu1)^-GPd_ML_ism$mle[2]-1)
  GPd_endpoint_ML_ism<-u1-GPd_ML_ism$mle[1]/GPd_ML_ism$mle[2]
    #evd
  GPd_ML_evd<-fpot(rpSeg,u1,model = 'gpd', start=list(scale=59,shape=-0.1))
  par(mfrow=c(2,2))
  plot(GPd_ML_evd)
  par(mfrow=c(1,1))
  GPd_prob542_ML_evd<-(Nu1/m)*(1+ GPd_ML_evd$estimate[2]*(542-u1)/
        GPd_ML_evd$estimate[1])^(-1/GPd_ML_evd$estimate[2])
  GPd_quant0.0001_ML_evd<-u1+GPd_ML_evd$estimate[1]/GPd_ML_evd$estimate[2]*
        ((m*0.0001/Nu1)^-GPd_ML_evd$estimate[2]-1)
  GPd_quant0.01_ML_evd<-u1+GPd_ML_evd$estimate[1]/GPd_ML_evd$estimate[2]*
```

```
        ((m*0.01/Nu1)^-GPd_ML_evd$estimate[2]-1)
  GPd_endpoint_ML_evd<-u1-GPd_ML_evd$estimate[1]/GPd_ML_evd$estimate[2]

#PWM - evir
  GPd_PWM_evir<-gpd(rpSeg,u1,method = 'pwm')
  sigmaevir<-as.numeric(GPd_PWM_evir$par.ests[2])
  xievir<-as.numeric(GPd_PWM_evir$par.ests[1])
  GPd_prob542_PWM_evir<-(Nu1/m)*(1+xievir*(542-u1)/sigmaevir)^(-1/xievir)
  GPd_quant0.0001_PWM_evir<-u1+GPd_PWM_evir$par.ests[2]/GPd_PWM_evir$par.ests[1]*
        ((m*0.0001/Nu1)^-GPd_PWM_evir$par.ests[1]-1)
  GPd_quant0.01_PWM_evir<-u1+GPd_PWM_evir$par.ests[2]/GPd_PWM_evir$par.ests[1]*
        ((m*0.01/Nu1)^-GPd_PWM_evir$par.ests[1]-1)
  GPd_endpoint_PWM_evir<-u1-GPd_PWM_evir$par.ests[2]/GPd_PWM_evir$par.ests[1]

#Confidence Intervals
    #parameters
  GPd_ML_CI<-confint(profile(GPd_ML_evd, conf = 0.95))
  GPd_ML_CI_scale<-c(GPd_ML_CI[1],GPd_ML_CI[3])
  GPd_ML_CI_shape<-c(GPd_ML_CI[2],GPd_ML_CI[4])
  par(mfrow=c(1,2))
  plot(profile(GPd_ML_evd, which='scale', conf=0.95))
  abline(v=GPd_ML_CI_scale[1],lty=2,col='grey50')
  abline(v=GPd_ML_CI_scale[2],lty=2,col='grey50')
  abline(v=GPd_ML_evd$estimate[1],lty=2,col='grey80')
  text(55,-2536.7,round(GPd_ML_CI_scale[1],4),col='grey50')
  text(64,-2536.7,round(GPd_ML_CI_scale[2],4),col='grey50')
  text(GPd_ML_evd$estimate[1],-2536.7,round(GPd_ML_evd$estimate[1],4),col='grey80')
  plot(profile(GPd_ML_evd,which='shape', conf=0.95))
  abline(v=GPd_ML_CI_shape[1],lty=2,col='grey50')
  abline(v=GPd_ML_CI_shape[2],lty=2,col='grey50')
  abline(v=GPd_ML_evd$estimate[2],lty=2,col='grey80')
  text(-0.19,-2536.7,round(GPd_ML_CI_shape[1],4),col='grey50')
  text(-0.11,-2536.7,round(GPd_ML_CI_shape[2],4),col='grey50')
  text(GPd_ML_evd$estimate[2],-2536.7,round(GPd_ML_evd$estimate[2],4),col='grey80')
  par(mfrow=c(1,1))

    #q(0.0001) and U(100)
  gpd.prof(GPd_ML_ism, m=10000,xlow=490,xup=580,npy = 1, nint=10000)
  abline(v=496.75,lty=2,col='grey50')
  abline(v=566.55,lty=2,col='grey50')
  abline(v=GPd_quant0.0001_ML_ism,lty=2,col='grey80')
  text(501,-2536.7,496.75,col='grey50')
```

```
  text(562,-2536.7,566.55,col='grey50')
  text(GPd_quant0.0001_ML_ism,-2536.7,round(GPd_quant0.0001_ML_ism,4),col='grey80')
  gpd.prof(GPd_ML_ism, m=100,xlow=405,xup=440,npy = 1, nint=10000)
  abline(v=408.59,lty=2,col='grey50')
  abline(v=435.65,lty=2,col='grey50')
  abline(v=GPd_quant0.01_ML_ism,lty=2,col='grey80')
  text(410.5,-2536.7,408.59,col='grey50')
  text(434,-2536.7,435.65,col='grey50')
  text(GPd_quant0.01_ML_ism,-2536.7,round(GPd_quant0.01_ML_ism,4),col='grey80')
```

A.19 Yearly Observations Plot

```
anos<-sort(unique(ANO)) ; maxleng<-0
for(i in 1:13){
  if(maxleng < length(rpSeg[ANO==2001+i]))
    maxleng<-length(rpSeg[ANO==2001+i])}
numb_obs<-c()
for(i in 1:13){numb_obs[i]<-length(rpSeg[ANO==2001+i])}
dados<-matrix(rep(NA,times=13*maxleng),13,maxleng)
for(i in 1:13){
  for(j in 1: numb_obs[i]){dados[i,j]<-sort(rpSeg[ANO==2001+i],decreasing = TRUE)[j]}}
dados<-cbind(anos,dados)
plot(anos,dados[,2],ylim=c(min(rpSeg),max(rpSeg)),type='p',xaxt='n',main='',
     ylab='Time (Sec)',xlab='Number of Records Set / Year')
axis(1,at=anos, labels=T, las=3)
axis(3,at=anos, labels=numb_obs, las=1.5, cex.axis=0.6)
for(i in 2:maxleng){points(anos,dados[,i])}
```

A.20 Largest 1, 5, 10 and 20 Yearly Observations Plots

```
par(mfrow=c(2,2))
plot(anos,dados[,2],type='p',xaxt='n',main='', ylim=c(170,550),
     ylab='Time (Sec)',xlab='Number of Records Set / Year')
axis(1,at=anos, labels=T, las=3)
axis(3,at=anos, labels=rep(1,13), las=1.5, cex.axis=0.6)
plot(anos,dados[,2],type='p',xaxt='n',main='', ylim=c(170,550),
     ylab='Time (Sec)',xlab='Number of Records Set / Year')
axis(1,at=anos, labels=T, las=3)
axis(3,at=anos, labels=rep(5,13), las=1.5, cex.axis=0.6)
for(i in 3:6){points(anos,dados[,i])}
plot(anos,dados[,2],type='p',xaxt='n',main='', ylim=c(170,550),
```

```
     ylab='Time (Sec)',xlab='Number of Records Set / Year')
axis(1,at=anos, labels=T, las=3)
axis(3,at=anos, labels=rep(10,13), las=1.5, cex.axis=0.6)
for(i in 3:11){points(anos,dados[,i])}
plot(anos,dados[,2],type='p',xaxt='n',main='', ylim=c(170,550),
     ylab='Time (Sec)',xlab='Number of Records Set / Year')
axis(1,at=anos, labels=T, las=3)
axis(3,at=anos, labels=rep(20,13), las=1.5, cex.axis=0.6)
for(i in 3:21){points(anos,dados[,i])}
par(mfrow=c(1,1))
```

A.21 Largest 1, 5, 10 and 20 Yearly Observations Fitting

```
rfit1<-rlarg.fit(dados[,-1],r=1)
rfit5<-rlarg.fit(dados[,-1],r=5)
rfit10<-rlarg.fit(dados[,-1],r=10)
rfit20<-rlarg.fit(dados[,-1],r=20)
rlarg.diag(rfit1)
rlarg.diag(rfit5)
rlarg.diag(rfit10)
rlarg.diag(rfit20)

mmo_prob542_r1<-1-pgev(542,rfit1$mle[1],rfit1$mle[2],rfit1$mle[3])
mmo_quant0.0001_r1<-qgev(0.9999,rfit1$mle[1],rfit1$mle[2],rfit1$mle[3])
mmo_quant0.01_r1<-qgev(0.99,rfit1$mle[1],rfit1$mle[2],rfit1$mle[3])
mmo_endpoint_r1<-rfit1$mle[1]-rfit1$mle[2]/rfit1$mle[3]
mmo_prob542_r5<-1-pgev(542,rfit5$mle[1],rfit5$mle[2],rfit5$mle[3])
mmo_quant0.0001_r5<-qgev(0.9999,rfit5$mle[1],rfit5$mle[2],rfit5$mle[3])
mmo_quant0.01_r5<-qgev(0.99,rfit5$mle[1],rfit5$mle[2],rfit5$mle[3])
mmo_endpoint_r5<-rfit5$mle[1]-rfit5$mle[2]/rfit5$mle[3]
mmo_prob542_r10<-1-pgev(542,rfit10$mle[1],rfit10$mle[2],rfit10$mle[3])
mmo_quant0.0001_r10<-qgev(0.9999,rfit10$mle[1],rfit10$mle[2],rfit10$mle[3])
mmo_quant0.01_r10<-qgev(0.99,rfit10$mle[1],rfit10$mle[2],rfit10$mle[3])
mmo_endpoint_r10<-rfit10$mle[1]-rfit10$mle[2]/rfit10$mle[3]
mmo_prob542_r20<-1-pgev(542,rfit20$mle[1],rfit20$mle[2],rfit20$mle[3])
mmo_quant0.0001_r20<-qgev(0.9999,rfit20$mle[1],rfit20$mle[2],rfit20$mle[3])
mmo_quant0.01_r20<-qgev(0.99,rfit20$mle[1],rfit20$mle[2],rfit20$mle[3])
mmo_endpoint_r20<-rfit20$mle[1]-rfit20$mle[2]/rfit20$mle[3]
```

A.22 Ratio, Hasofer-Wang and Greenwood Tests

```
m<-length(name); K<-1:(m-1); XX<-sort(rpSeg)   #ordered sample

Njk<-matrix(seq(1:(2*(m-1))),2,m-1)
for(j in 1:2){
  for(k in K){ Njk[j,k]<-(1/k)*sum((XX[(m-k+1):m]-XX[m-k])^j)}}
#Greenwood Statistic
Gr<-c() ; Gr_ast<-c()
for(k in K){
  Gr[k]<-Njk[2,k]/(Njk[1,k]^2)
  Gr_ast[k]<-sqrt(k/4)*(Gr[k]-2)}
#Hasofer-Wang Statistic
W<-c() ; W_ast<-c()
for(k in K){
  W[k]<-(1/k)*(1/(Gr[k]-1))
  W_ast[k]<-sqrt(k/4)*(k*W[k]-1)}
#Ratio Statistic
R<-c() ; R_ast<-c()
for(k in K){
  R[k]<-(XX[m]-XX[m-k])/Njk[1,k]
  R_ast[k]<-R[k]-log(k)}

#Plot - two-sided test
plot(K,R_ast,type='n', ylim=c(-8,20), xlab='k',ylab='Observed Statistics')
points(K,Gr_ast, type='l', col='grey20')
points(K,W_ast, type='l', col='grey50')
points(K,R_ast, type='l', col='grey80')
legend(0,20,c('Greenwood','Hasofer-Wang','Ratio'), seg.len=1,
       lty=c(1,1,1), lwd=c(1,1,1), col=c('grey20','grey50','grey80'))
abline(h=qnorm(0.05/2), lty=2, lwd=2)
abline(h=qnorm(1-0.05/2), lty=2, lwd=2)
abline(h=qgumbel(0.05/2), lty=3)
abline(h=qgumbel(1-0.05/2), lty=3)
legend(0,12,c(expression(list(z[0.025],z[0.975])),expression(list(g[0.025],g[0.975]))),
        seg.len=2,lty=c(2,3), lwd=c(2,1))
#Plot - one-sided test EVI<0
plot(K,R_ast,type='n', ylim=c(-8,20), xlab='k',ylab='Observed Statistics')
points(K,Gr_ast, type='l', col='grey20')
points(K,W_ast, type='l', col='grey50')
points(K,R_ast, type='l', col='grey80')
legend(0,20,c('Greenwood','Hasofer-Wang','Ratio'), seg.len=1,
```

```
       lty=c(1,1,1), lwd=c(1,1,1), col=c('grey20','grey50','grey80'))
abline(h=qnorm(0.05), lty=2, lwd=2)
abline(h=qnorm(1-0.05), lty=6, lwd=2)
abline(h=qgumbel(0.05), lty=3, lwd=2)
legend(0,12,c(expression(z[0.05]),expression(z[0.95]), expression(g[0.05])),seg.len=2,
       lty=c(2,6,3), lwd=c(2,2,2))
```

A.23 General Right Endpoint Estimator Sample Path

```
K2<-1:(m/2) ; gen_xF<-c()
for(k in K2){
  gen_xF[k]<-XX[m]+XX[m-k]-(1/log(2))*sum(log(1+1/(k+(0:(k-1))))*XX[(m-k):(m-k-k+1)])}
plot(2*K2,gen_xF, type='l', xlab='k=2k*', ylab=expression(hat(x)[k]*paste('')^F*
        paste(' (Sec)')),col='grey50',mgp=c(2.5,1,0))
abline(h=max(XX), lty=2)
text(390,543,expression(X[n:n]==542))
legend(0,580,c(expression(paste('General Right Endpoint Estimator '*
        hat(x)[k]*paste('')^F))),lty=c(1), col=c('grey50'))
```

A.24 Test Statistic based on the General Right Endpoint Estimator

```
Gnk<-c() ; Gnk0_ast<-c()
for(k in K2){
  Gnk[k]<-(gen_xF[k]-XX[m-k])/(XX[m-k]-XX[m-2*k])
  Gnk0_ast[k]<-log(2)*Gnk[k]-(log(k)+log(2)/2)}
plot(2*K2,Gnk0_ast,type='l', xlab='k',ylab='Observed Statistics', col='grey50')
legend(630,12,c(expression(G[list(m,k)]^paste('*')*paste('(0)'))), seg.len=2,
       lty=c(1), lwd=c(1), col=c('grey50'))
abline(h=qgumbel(0.05/2), lty=3)
abline(h=qgumbel(1-0.05/2), lty=3)
abline(h=qgumbel(0.05), lty=6)
legend(630,10,c(expression(list(g[0.025],g[0.975])),expression(g[0.05])),seg.len=2,
       lty=c(3,6), lwd=c(1,1))
```

A.25 Finiteness of the Right Endpoint Test Sample Path

```
logNjk<-matrix(seq(1:(2*(m-1))),2,m-1)
for(j in 1:2){ for(k in K){
    logNjk[j,k]<-(1/k)*sum((log(XX[(m-k+1):m])-log(XX[m-k]))^j)}}
```

```
TT<-c(); TT1<-c(); TT1_ast<-c()
for(k in K){
  TT[k]<-XX[m-k]*logNjk[1,k]*0.5*(1-(logNjk[1,k]^2)/logNjk[2,k])^-1
  TT1[k]<-(1/k)*(sum((XX[(m-k):(m-1)]-XX[m-k]-TT[k])/(XX[m]-XX[m-k])))
  TT1_ast[k]<-sqrt(k)*log(k)*TT1[k]}
plot(K,TT1_ast, type='l',ylim=c(-32,2),xlab='k',ylab='Observed Statistics',
       col='grey50')
abline(h=qnorm(0.05/2), lty=2, lwd=2)
abline(h=qnorm(1-0.05/2), lty=2, lwd=2)
legend(0,-25,c(expression(T[1]^paste('*'))), seg.len=2,
       lty=c(1), lwd=c(1), col=c('grey50'))
legend(0,-28,c(expression(list(z[0.025],z[0.975]))),seg.len=2,
       lty=c(2), lwd=c(2))
```

A.26 Pickands Estimator

```
xi_pic<-c()
for(k in 1:(m/4)){xi_pic[k]<-(1/log(2))*log((XX[m-k+1]-XX[m-2*k+1])/
        (XX[m-2*k+1]-XX[m-4*k+1]))}
plot(4*(1:(m/4)),xi_pic, type='l',xlab='k=4K*',
        ylab=expression(hat(xi)[list(m,k)]^P),col='grey30',mgp=c(2.2,1,0))
abline(h=0,lty=2,col='black')
abline(h=-0.16,lty=3,col='grey60')
text(765,-0.3, expression(hat(xi)['POT240']==-0.16), col='grey60')
```

A.27 Generalized Hill Estimator

```
K<-1:(m-1) ; Mjk<-matrix(seq(1:(2*(m-1))),2,m-1)
for(j in 1:2){
  for(k in K){ soma<-0
    for(i in 1:(k)){ soma<-soma+(log(XX[m-i+1])-log(XX[m-k]))^j}
    Mjk[j,k]<-(1/k)*soma}}
xi_ghill<-c()
for(k in K){ soma<-0
  for(i in 1:k){ soma<-soma+(log(Mjk[1,i])-log(Mjk[1,k]))}
  xi_ghill[k]<-Mjk[1,k]+(1/k)*soma}
plot((K+1),xi_ghill, type='l',xlab='k+1',ylab=expression(hat(xi)[list(m,k)]^GH),
       col='grey30',mgp=c(2.2,1,0))
abline(h=0,lty=2,col='black')
abline(h=-0.16,lty=3,col='grey60')
text(765,-0.21, expression(hat(xi)['POT240']==-0.16), col='grey60')
```

A.28 Moment and Negative Moment Estimators

```
xi_Mom_pos<-c() ; xi_Mom_neg<-c() ; xi_Mom<-c()
for(k in K){
  xi_Mom_pos[k]<-Mjk[1,k]
  xi_Mom_neg[k]<-1-(1/2)*(1-(Mjk[1,k]^2)/Mjk[2,k])^-1
  xi_Mom[k]<-xi_Mom_pos[k]+xi_Mom_neg[k]}
plot((K+1),xi_Mom, type='l',xlab='k+1',ylab=expression(hat(xi)[list(m,k)]^list(M,NM)),
     col='grey50',mgp=c(2.2,1,0))
points((K+1),xi_Mom_neg, type='l',xlab='k',
        ylab=expression(hat(xi)[list(m,k)]^paste('-')), col='grey30',mgp=c(2.2,1,0))
legend(590,0.2,c('Moment','Negative Moment'), col=c('grey50','grey30'), lty=c(1,1))
abline(h=0,lty=2,col='black')
abline(h=-0.16,lty=3,col='grey60')
text(765,-0.23, expression(hat(xi)['POT240']==-0.16), col='grey60')
```

A.29 Mixed Moment Estimator

```
L<-c() ; fi<-c() ; xi_MixMom<-c()
for(k in K){ soma<-0
  for(i in 1:k){ soma<-soma+1-(XX[m-k]/XX[m-i+1])}
  L[k]<-(1/k)*soma
  fi[k]<-(Mjk[1,k]-L[k])/(L[k]^2)
  xi_MixMom[k]<-(fi[k]-1)/(1+2*min(fi[k]-1,0))}
plot((K+1)[-1],xi_MixMom[-1], type='l',xlab='k+1',ylab=expression(hat(xi)[list(m,k)]^MM),
     col='grey30',mgp=c(2.2,1,0))
abline(h=0,lty=2,col='black')
abline(h=-0.16,lty=3,col='grey60')
text(765,-0.12, expression(hat(xi)['POT240']==-0.16), col='grey60')
```

A.30 Location Invariant Moment Estimator

```
Njk<-matrix(seq(1:(2*(m-1))),2,m-1)
for(j in 1:2){
  for(k in K){ Njk[j,k]<-(1/k)*sum((XX[(m-k+1):m]-XX[m-k])^j)}}
xi_art<-c()
for(k in K){ xi_art[k]<-1-(1/2)*(1-(Njk[1,k]^2)/Njk[2,k])^-1}
plot((K+1),xi_art, type='l',xlab='k+1',
        ylab=expression(hat(xi)[list(m,k)]^IM),col='grey50',mgp=c(2.2,1,0))
abline(h=0,lty=2,col='black') ; abline(h=-0.16,lty=3,col='grey60')
text(765,-0.22, expression(hat(xi)['POT240']==-0.16), col='grey60')
```

A.31 PORT-Moment Estimator

```
Q<-c(0,0.1,0.2,0.5)
for(i in 1:4){
  q<-Q[i] ; nq<-floor(m*q)+1 ; Xq<-XX-XX[nq]
  nn<-m-nq-1 ; K<-1:(nn) ; xi_PORT_Mom<-c()
  for(k in K){
    y1<-NULL ; y2<-NULL ; M1<-NULL ; M2<-NULL
    for(j in 1:k) {
      y1[j]<-(log(Xq[m-j+1])-log(Xq[m-k]))^1
      y2[j]<-(log(Xq[m-j+1])-log(Xq[m-k]))^2}
    M1<-(1/k)*sum(y1) ; M2<-(1/k)*sum(y2)
    xi_PORT_Mom[k]<- M1+1-(1/2)*(1-(M1)^2/M2)^-1}
  if(i==1){
    plot(K+1,xi_PORT_Mom, type='l',xlab='k',
            ylab=expression(hat(xi)[list(m,k)]^M(q)),
            main='',col='black',mgp=c(2.2,1,0))
  }else if(i==2){
    points(K+1,xi_PORT_Mom, type='l',col='grey80',lwd=2)
  }else if(i==3){
    points(K+1,xi_PORT_Mom, type='l',col='black',lty=3,lwd=1.6)
  }else if(i==4){
    points(K+1,xi_PORT_Mom, type='l',col='grey50')}
  abline(h=0,lty=2,col='black')}
legend(15,-2.5,legend=c("q=0","q=0.1","q=0.2","q=0.5"),lty=c(1,1,3,1),
       col=c("black","grey80","black","grey50"),lwd=c(1,2,1.6,1))

q<-0.1 ; nq<-floor(m*q)+1 ; Xq<-XX-XX[nq]
NN<-m-nq-1 ; K<-1:(NN) ; xi_PORT_Mom<-c()
for(k in K){
  y1<-NULL ; y2<-NULL ; M1<-NULL ; M2<-NULL
  for(j in 1:k) {
    y1[j]<-(log(Xq[m-j+1])-log(Xq[m-k]))^1
    y2[j]<-(log(Xq[m-j+1])-log(Xq[m-k]))^2}
  M1<-(1/k)*sum(y1) ; M2<-(1/k)*sum(y2)
  xi_PORT_Mom[k]<- M1+1-(1/2)*(1-(M1)^2/M2)^-1}
```

A.32 PORT-Mixed Moment Estimator

```
Q<-c(0,0.01,0.1,0.2)
for(t in 1:4){
  q<-Q[t] ; nq<-floor(m*q)+1 ; Xq<-sort(XX[(nq+1):m]-XX[nq])
```

```
  nn<-length(Xq) ; K<-1:(nn-1)
  Mjk<-matrix(seq(1:(2*(nn-1))),2,nn-1)
  for(j in 1:2){
    for(k in K){ soma<-0
      for(i in 1:(k)){ soma<-soma+(log(Xq[nn-i+1])-log(Xq[nn-k]))^j}
      Mjk[j,k]<-(1/k)*soma}}
  L<-c() ; fi<-c() ; xi_PORT_MixMom<-c()
  for(k in K){ soma<-0
    for(i in 1:k){ soma<-soma+1-(Xq[nn-k]/Xq[nn-i+1])}
    L[k]<-(1/k)*soma ; fi[k]<-(Mjk[1,k]-L[k])/(L[k]^2)
    xi_PORT_MixMom[k]<-(fi[k]-1)/(1+2*min(fi[k]-1,0))}
  if(t==1){
    plot(K+1,xi_PORT_MixMom, type='l',xlab='k',
            ylab=expression(hat(xi)[list(m,k)]^MM(q)),
            main='',col='black',mgp=c(2.2,1,0))
  }else if(t==2){
    points(K+1,xi_PORT_MixMom, type='l',col='grey80', lwd=2)
  }else if(t==3){
    points(K+1,xi_PORT_MixMom, type='l',col='black',lty=3,lwd=1.6)
  }else if(t==4){
    points(K+1,xi_PORT_MixMom, type='l',col='grey50')}}
legend(15,2.3,legend=c("q=0","q=0.01","q=0.1","q=0.2"),lty=c(1,1,3,1),
       col=c("black","grey80","black","grey50"),lwd=c(1,2,1.6,1))
abline(h=0,lty=2,col='black')
```

A.33 All Estimators Plot

```
K<-(1:(m-1))
plot((4*(1:(m/4)))[-1],xi_pic[-1], type='l',xlab='k+1',
    ylab=expression(hat(xi)[list(m,k)]^T),mgp=c(2.2,1,0)) #Pickands
points(K+1,xi_Mom, type='l', col='grey60',lty=3)   #Moment
points(K+1,xi_Mom_neg, type='l',col='grey40',lty=4) #Negative Moment
points(K+1,xi_ghill, type='l', col='black',lty=5)  #Generalized Hill
points(K+1,xi_MixMom, type='l', col='grey80', lwd=2) #Mixed Moment
points(K+1,xi_art, type='l', col='black', lwd=2)  #MOM.inv
points((1:NN)+1, xi_PORT_Mom, type='l',col='lightslategray') #PORT-Moment q=0.1
abline(h=0,lty=2,col='grey80')
legend(250,1.2,c('Pickands','Moment','Negative Moment', "Generalized Hill",
        'Mixed-Moment','Invariant Moment' ,'PORT-Moment q=0.1'),
        seg.len=c(2,2,2,2,2,2,2), lty=c(1,3,4,5,1,1,1), lwd=c(1,1,1,1,2,2,1),
        col=c('black','grey60','grey40','black','grey80','black','lightslategray'))
text(780,0.05, expression(hat(xi)==0), col='grey60')
```

A.34 Heuristic Choice of Tail Sample Fraction and Plot

```
PICK<-xi_pic
MOM<-xi_Mom[4*(1:(m/4))]
MOMNEG<-xi_Mom_neg[4*(1:(m/4))]
GHILL<-xi_ghill[4*(1:(m/4))]
MIXMOM<-xi_MixMom[4*(1:(m/4))]
MOMINV<-xi_art[4*(1:(m/4))]
PORTMOM<-xi_PORT_Mom[4*(1:(m/4))]
difquad<-(PICK-MOM)^2+(PICK-MOMNEG)^2+(PICK-GHILL)^2+(PICK-MIXMOM)^2+
    (PICK-PORTMOM)^2+(PICK-MOMINV)^2+(MOM-MOMNEG)^2+(MOM-GHILL)^2+(MOM-MIXMOM)^2+
    (MOM-PORTMOM)^2+(MOM-MOMINV)^2+(MOMNEG-GHILL)^2+(MOMNEG-MIXMOM)^2+
    (MOMNEG-PORTMOM)^2+(MOMNEG-MOMINV)^2+(GHILL-MIXMOM)^2+(GHILL-PORTMOM)^2+
    (GHILL-MOMINV)^2+(MIXMOM-PORTMOM)^2+(MIXMOM-MOMINV)^2+(PORTMOM-MOMINV)^2
kot<-which.min(difquad)*4

plot(4*(1:(m/4)), difquad, type='l', xlab='k',
     ylab=expression(sum((hat(xi)[list(m,k)]^(i)-hat(xi)[list(m,k)]^(j))^2)),
     mgp=c(2.2,1,0))
abline(h=0,lty=2,col='grey50')
abline(v=kot, lty=3, col='slategrey')
text(260,40,expression(k^paste(opt)==216), col='slategrey')
```

A.35 POT-ML Estimator and Sample Path Plot

```
est<-matrix(0, nrow = m-2, ncol = 2) ; u<-NULL
for(k in 2:(m-1)){
  u[k-1]<-XX[m-k]
  aux<-gpd.fit(rpSeg,u[k-1])
  est[k-1,]<-aux$mle}
K<-(1:(m-1))
plot(K+1,xi_ghill, type='l',xlab='k+1',ylab=expression(hat(xi)[list(m,k)]^T),
        mgp=c(2.2,1,0)) #Gen Hill
points(K+1,xi_MixMom, type='l', col='grey60', lwd=2) #Mix Moment
points(K+1,xi_art, type='l', col='black', lwd=1, lty=2)  #MOM.inv artigo
points(2:(m-1),est[,2], type='l',col='lightslategray',lwd=2) #POT-ML
abline(h=0,lty=2,col='grey80')
legend(250,0.5,c('Generalized Hill','Mixed-Moment','Invariant Moment' ,'POT-ML'),
       seg.len=c(2,2,2,2), lty=c(1,1,2,1), lwd=c(1,2,1,2),pch=c(8,15,19,25),
       col=c('black','grey60','black','lightslategray'))
text(780,0.05, expression(hat(xi)==0), col='grey60')
abline(v=kot+1, lty=2, lwd=1, col='slategrey')
```

```
text(250,-0.4,expression(k^paste(opt)==216), col='slategrey')
points(217,xi_MixMom[216], pch=15,col='grey60', cex=1.5)
points(217,xi_art[216],pch=19,cex=1.5)
points(217,est[215,2],pch=25,col='slategrey', cex=1.5,lwd=2)
points(217,xi_ghill[216], pch=8, cex=1.2)
```

A.36 Semi-parametric EVI Estimation

```
xi_pic_216<-xi_pic[54]
xi_Mom_216<-xi_Mom[216]
xi_Mom_neg_216<-xi_Mom_neg[216]
xi_ghill_216<-xi_ghill[216]
xi_MixMom_216<-xi_MixMom[216]
xi_art_216<-xi_art[216]
xi_PORT_Mom_216<-xi_PORT_Mom[216]
xi_pot_ml_216<-gpd.fit(rpSeg,XX[m-216])$mle[2]

#95% CI's
interconf<-function(estimador,variass){
  IC<-c()
  IC[1]<-estimador-qnorm(0.975)*sqrt(variass/216)
  IC[2]<-estimador+qnorm(0.975)*sqrt(variass/216)
  return(IC)}
    #Pickands
var_pic<- function(xi){
  return(xi^2*(2^(2*xi+1)+1)/(((2^xi-1)*log(2))^2))}
var_pic_216<-var_pic(xi_pic_216)
IC_pic_216<-interconf(xi_pic_216,var_pic_216)
    #Generalized Hill
var_ghill<- function(xi){
  var<-xi
  xi1<-xi[xi<0]; xi2<-xi[xi>=0]
  var[xi<0]<-(1-xi1)*(1+xi1+2*xi1^2)/(1-2*xi1)
  var[xi>=0]<-1+xi2^2
  return(var)}
var_ghill_216<-var_ghill(xi_ghill_216)
IC_ghill_216<-interconf(xi_ghill_216,var_ghill_216)
    #Moment, Negative Moment and PORT-Moment q=0.1
var_Mom<- function(xi) {
  var<-xi
  xi1<-xi[xi<0]; xi2<-xi[xi>=0]
  var[xi<0]<-(1-xi1)^2*(1-2*xi1)*(4-8*(1-2*xi1)/(1-3*xi1)+
```

```
        (5-11*xi1)*(1-2*xi1)/((1-3*xi1)*(1-4*xi1)))
  var[xi>=0]<-1+xi2^2
  return(var)}
var_Mom_216<-var_Mom(xi_Mom_216)
IC_Mom_216<-interconf(xi_Mom_216,var_Mom_216)
var_Mom_neg_216<-var_Mom(xi_Mom_neg_216)
IC_Mom_neg_216<-interconf(xi_Mom_neg_216,var_Mom_neg_216)
var_PORT_Mom_216<-var_Mom(xi_PORT_Mom_216)
IC_PORT_Mom_216<-interconf(xi_PORT_Mom_216,var_PORT_Mom_216)
    #Mixed-Moment
var_MixMom<- function(xi) {
  var<-xi
  xi1<-xi[xi<0]; xi2<-xi[xi>=0]
  var[xi<0]<-((1-2*xi1)^4)*(((1-xi1)^2*(6*xi1^2-xi1+1))/
        ((1-2*xi1)^3*(1-3*xi1)*(1-4*xi1)))
  var[xi>=0]<-(1+xi2)^2
  return(var)}
var_MixMom_216<-var_MixMom(xi_MixMom_216)
IC_MixMom_216<-interconf(xi_MixMom_216,var_MixMom_216)
    #POT-ML
var_pot_ml<-function(xi) {
  var<-(xi+1)^2
  return(var)}
var_pot_ml_216<-var_pot_ml(xi_pot_ml_216)
IC_pot_ml_216<-interconf(xi_pot_ml_216,var_pot_ml_216)
```

A.37 Location and Scale Attraction Coefficients Estimation

```
K<-1:(m-1) ; Mjk<-matrix(seq(1:(2*(m-1))),2,m-1)
for(j in 1:2){
  for(k in K){ soma<-0
    for(i in 1:(k)){ soma<-soma+(log(XX[m-i+1])-log(XX[m-k]))^j}
    Mjk[j,k]<-(1/k)*soma}}
b_mk<-XX[m-216] ; a_mk<-XX[m-216]*Mjk[1,216]*(1-xi_Mom_neg[216])
a_mk_LOCINV<-Njk[1,216]*(1-xi_art_216)
```

A.38 Semi-parametric Estimation of Indicators of Interest

```
prob542_pic_216<-(216/m)*(1-evd::pgpd(542-b_mk,scale=a_mk, shape=xi_pic_216))
prob542_Mom_216<-(216/m)*(1-evd::pgpd(542-b_mk,scale=a_mk, shape=xi_Mom_216 ))
prob542_Mom_neg_216<-(216/m)*(1-evd::pgpd(542-b_mk,scale=a_mk,
```

```
        shape=xi_Mom_neg_216))
prob542_ghill_216<-(216/m)*(1-evd::pgpd(542-b_mk,scale=a_mk, shape=xi_ghill_216))
prob542_MixMom_216<-(216/m)*(1-evd::pgpd(542-b_mk,scale=a_mk, shape=xi_MixMom_216))
prob542_MOMINV_216<-(216/m)*(1-evd::pgpd(542-b_mk,scale=a_mk_LOCINV,shape=xi_art_216))
prob542_PORT_Mom_216<-(216/m)*(1-evd::pgpd(542-b_mk,scale=a_mk,
        shape=xi_PORT_Mom_216))
prob542_pot_ml_216<-(216/m)*(1-evd::pgpd(542-b_mk,scale=a_mk, shape=xi_pot_ml_216))

quant_0.9999_pic_216<-XX[m-216]+a_mk*((216/(m*0.0001))^xi_pic_216-1)/xi_pic_216
quant_0.9999_Mom_216<-XX[m-216]+a_mk*((216/(m*0.0001))^xi_Mom_216-1)/xi_Mom_216
quant_0.9999_Mom_neg_216<-XX[m-216]+a_mk*((216/(m*0.0001))^xi_Mom_neg_216-1)/
        xi_Mom_neg_216
quant_0.9999_ghill_216<-XX[m-216]+a_mk*((216/(m*0.0001))^xi_ghill_216-1)/xi_ghill_216
quant_0.9999_MixMom_216<-XX[m-216]+a_mk*((216/(m*0.0001))^xi_MixMom_216-1)/xi_MixMom_216
quant_0.9999_MOMINV_216<-XX[m-216]+a_mk_LOCINV*((216/(m*0.0001))^xi_art_216-1)/xi_art_216
quant_0.9999_PORT_Mom_216<-XX[m-216]+a_mk*((216/(m*0.0001))^xi_PORT_Mom_216-1)/
        xi_PORT_Mom_216
quant_0.9999_pot_ml_216<-XX[m-216]+a_mk*((216/(m*0.0001))^xi_pot_ml_216-1)/xi_pot_ml_216

quant_0.99_pic_216<-XX[m-216]+a_mk*((216/(m*0.01))^xi_pic_216-1)/xi_pic_216
quant_0.99_Mom_216<-XX[m-216]+a_mk*((216/(m*0.01))^xi_Mom_216-1)/xi_Mom_216
quant_0.99_Mom_neg_216<-XX[m-216]+a_mk*((216/(m*0.01))^xi_Mom_neg_216-1)/
        xi_Mom_neg_216
quant_0.99_ghill_216<-XX[m-216]+a_mk*((216/(m*0.01))^xi_ghill_216-1)/xi_ghill_216
quant_0.99_MixMom_216<-XX[m-216]+a_mk*((216/(m*0.01))^xi_MixMom_216-1)/xi_MixMom_216
quant_0.99_MOMINV_216<-XX[m-216]+a_mk_LOCINV*((216/(m*0.01))^xi_art_216-1)/xi_art_216
quant_0.99_PORT_Mom_216<-XX[m-216]+a_mk*((216/(m*0.01))^xi_PORT_Mom_216-1)/
        xi_PORT_Mom_216
quant_0.99_pot_ml_216<-XX[m-216]+a_mk*((216/(m*0.01))^xi_pot_ml_216-1)/xi_pot_ml_216

endpoint_pic_216<-XX[m-216]*(1-Mjk[1,216]*(1-xi_Mom_neg[216])/xi_pic_216)
endpoint_Mom_216<-XX[m-216]*(1-Mjk[1,216]*(1-xi_Mom_neg[216])/xi_Mom_216)
endpoint_Mom_neg_216<-XX[m-216]*(1-Mjk[1,216]*(1-xi_Mom_neg[216])/xi_Mom_neg_216)
endpoint_ghill_216<-XX[m-216]*(1-Mjk[1,216]*(1-xi_Mom_neg[216])/xi_ghill_216)
endpoint_MixMom_216<-XX[m-216]*(1-Mjk[1,216]*(1-xi_Mom_neg[216])/xi_MixMom_216)
endpoint_MOMINV_216<-XX[m-216] - (a_mk_LOCINV/xi_art_216)
endpoint_PORT_Mom_216<-XX[m-216]*(1-Mjk[1,216]*(1-xi_Mom_neg[216])/xi_PORT_Mom_216)
endpoint_pot_ml_216<-XX[m-216]*(1-Mjk[1,216]*(1-xi_Mom_neg[216])/xi_pot_ml_216)
endpoint_gen_xF_216<-gen_xF[108]
```

A.39 Right Endpoint Estimators' Sample Paths with k^{opt}

```
#COMPLETE
endpoint_Mom_k<-c()
endpoint_Mom_neg_k<-c()
endpoint_ghill_k<-c()
endpoint_MixMom_k<-c()
endpoint_pic_k<-c()
endpoint_MOMINV_k<-c()
endpoint_PORT_Mom_k<-c()
endpoint_pot_ml_k<-c(NaN)
for(k in K){
  endpoint_Mom_k[k]<-max(XX[m],XX[m-k]*(1-Mjk[1,k]*(1-xi_Mom_neg[k])/xi_Mom[k]))
  endpoint_Mom_neg_k[k]<-max(XX[m],XX[m-k]*(1-Mjk[1,k]*(1-xi_Mom_neg[k])/
        xi_Mom_neg[k]))
  endpoint_ghill_k[k]<-max(XX[m],XX[m-k]*(1-Mjk[1,k]*(1-xi_Mom_neg[k])/xi_ghill[k]))
  endpoint_MixMom_k[k]<-max(XX[m],XX[m-k]*(1-Mjk[1,k]*(1-xi_Mom_neg[k])/xi_MixMom[k]))
  endpoint_MOMINV_k[k]<-max(XX[m],XX[m-k]-Njk[1,k]*(1-xi_art[k])/xi_art[k])}
for(k in 2:(m-1)){
  endpoint_pot_ml_k[k]<-max(XX[m],XX[m-k]*(1-Mjk[1,k]*(1-xi_Mom_neg[k])/est[k-1,2]))}
for(k in 1:(m/4)){
  endpoint_pic_k[k]<-max(XX[m],XX[m-4*k]*(1-Mjk[1,4*k]*(1-xi_Mom_neg[4*k])/xi_pic[k]))}
for(k in 1:NN){
  endpoint_PORT_Mom_k[k]<-max(XX[m],XX[m-k]*(1-Mjk[1,k]*(1-xi_Mom_neg[k])/
        xi_PORT_Mom[k]))}

plot(K+1,endpoint_ghill_k, type='l', xlab='k+1',ylim=c(540,2000),
     ylab=expression(hat(x)[k]*paste('')^F*paste(' (Sec)')),mgp=c(2.5,1,0))
points(K+1,endpoint_Mom_k, type='l', col='red')
points(K+1,endpoint_Mom_neg_k, type='l', col='blue')
points(K+1,endpoint_MixMom_k, type='l', col='green')
points(K+1,endpoint_MOMINV_k, type='l', col='purple')
points(K+1,endpoint_pot_ml_k, type='l', col='yellow')
points(1:NN+1,endpoint_PORT_Mom_k, type='l', col='cyan3')
points(2*K2+1,gen_xF, type='l', col='grey50')
points((4*(1:(m/4)))+1,endpoint_pic_k, type='l', col='orange' )
abline(h=max(XX), lwd=3)
abline(v=kot+1, lty=2)
text(780,570,expression(X[n:n]==542))
text(270,2000,expression(k^paste(opt)==216), col='slategrey')
legend(580,2000,c('General', 'Generalized Hill','Moment',
        'Negative Moment','Mixed-Moment', 'Pickands', 'MOM.inv',
```

```
        'PORT-Moment q=0.1', 'POT-ML'),lty=c(1,1,1,1,1,1,1,1,1),
        col=c('grey50','black','red','blue','green','orange','purple',
        'cyan3', 'yellow'))

#210<k<220
plot(K+1,endpoint_ghill_k, type='l', xlab='k+1', xlim=c(210,220),ylim=c(540,2000),
     ylab=expression(hat(x)[k]*paste('')^F*paste(' (Sec)')),mgp=c(2.5,1,0))
points(K+1,endpoint_Mom_k, type='l', col='red')
points(K+1,endpoint_Mom_neg_k, type='l', col='blue')
points(K+1,endpoint_MixMom_k, type='l', col='green')
points(K+1,endpoint_MOMINV_k, type='l', col='purple')
points(K+1,endpoint_pot_ml_k, type='l', col='yellow')
points(1:NN+1,endpoint_PORT_Mom_k, type='l', col='cyan3')
points(2*K2+1,gen_xF, type='l', col='grey50')
points((4*(1:(m/4)))+1,endpoint_pic_k, type='l', col='orange' )
abline(h=max(XX), lwd=3)
abline(v=217, lty=2)
text(210.3,510,expression(X[n:n]==542))
text(217.5,650,expression(k^paste(opt)==216), col='slategrey')
legend(217.38,2060,c('General', 'Generalized Hill','Moment','Negative Moment',
        'Mixed-Moment', 'Pickands', 'MOM.inv', 'PORT-Moment q=0.1', 'POT-ML'),
       lty=c(1,1,1,1,1,1,1,1,1), col=c('grey50','black','red','blue','green',
       'orange','purple', 'cyan3', 'yellow'),text.font=c(2,2,2,1,2,2,2,2,2))
```

A.40 Right Endpoint Choice of Tail Sample Fraction Heuristic

```
ENDMOM<-endpoint_Mom_k[2*(1:(NN/2))]
ENDMOMNEG<-endpoint_Mom_neg_k[2*(1:(NN/2))]
ENDGHILL<-endpoint_ghill_k[2*(1:(NN/2))]
ENDMIXMOM<-endpoint_MixMom_k[2*(1:(NN/2))]
ENDMOMINV<-endpoint_MOMINV_k[2*(1:(NN/2))]
ENDPORTMOM<-endpoint_PORT_Mom_k[2*(1:(NN/2))]
ENDGEN<-gen_xF[(1:(NN/2))]

difquad<-(ENDMOM-ENDMOMNEG)^2+(ENDMOM-ENDGHILL)^2+(ENDMOM-ENDMIXMOM)^2+
        (ENDMOM-ENDPORTMOM)^2+(ENDMOM-ENDMOMINV)^2+(ENDMOMNEG-ENDGHILL)^2+
        (ENDMOMNEG-ENDMIXMOM)^2+(ENDMOMNEG-ENDPORTMOM)^2+(ENDMOMNEG-ENDMOMINV)^2+
        (ENDGHILL-ENDMIXMOM)^2+(ENDGHILL-ENDPORTMOM)^2+(ENDGHILL-ENDMOMINV)^2+
        (ENDMIXMOM-ENDPORTMOM)^2+(ENDMIXMOM-ENDMOMINV)^2+(ENDPORTMOM-ENDMOMINV)^2+
        (ENDMOM-ENDGEN)^2+(ENDMOMNEG-ENDGEN)^2+(ENDGHILL-ENDGEN)^2+
        (ENDMIXMOM-ENDGEN)^2+(ENDMOMINV-ENDGEN)^2+(ENDPORTMOM-ENDGEN)^2
kot<-which.min(difquad)
```

```
kot6<-which.min(difquad[50:100])+49
kot7<-which.min(difquad[100:150])+99
kot8<-which.min(difquad[150:200])+149

plot(2*(1:(NN/2))+1, difquad, type='l', xlab='k', ylim=c(0,600000),
     ylab=expression(sum((hat(xi)[list(m,k)]^(i)-
     hat(xi)[list(m,k)]^(j))^2)),mgp=c(2.2,1,0))
abline(h=0,lty=2,col='grey80')
abline(v=217, lty=3, col='slategrey')
points(129,difquad[64],cex=1.5, lwd=2)
points(267,difquad[133],cex=1.5, lwd=2)
points(371,difquad[185],cex=1.5, lwd=2)
abline(v=267, lty=3, col='slategrey')
abline(v=371, lty=3, col='slategrey')
abline(v=129, lty=3, col='slategrey')
text(150,-10,128, col='slategrey',cex=0.9)
text(230,-10,216, col='slategrey',cex=0.9)
text(285,-10,266, col='slategrey',cex=0.9)
text(395,-10,370, col='slategrey',cex=0.9)
```

A.41 Right Endpoint Estimators' Sample Paths – 3 Ranges of Stability

```
#125<k<132
plot(K+1,endpoint_ghill_k, type='l', xlab='k+1', xlim=c(125,132), ylim=c(540,1000),
     ylab=expression(hat(x)[k]*paste('')^F*paste(' (Sec)')),mgp=c(2.5,1,0))
points(K+1,endpoint_Mom_k, type='l', col='red')
points(K+1,endpoint_Mom_neg_k, type='l', col='blue')
points(K+1,endpoint_MixMom_k, type='l', col='green')
points(K+1,endpoint_MOMINV_k, type='l', col='purple')
points(K+1,endpoint_pot_ml_k, type='l', col='yellow')
points(1:NN+1,endpoint_PORT_Mom_k, type='l', col='cyan3')
points(2*K2+1,gen_xF, type='l', col='grey50')
abline(h=max(XX), lwd=3)
abline(v=129, lty=2, lwd=2.5, col='slategrey')
text(129.4,900,expression(k^paste(opt)==128), col='slategrey')
text(131.5,520,expression(X[n:n]==542))
legend(130,1000,c('General', 'Generalized Hill','Moment','Negative Moment',
        'Mixed-Moment', 'MOM.inv', 'PORT-Moment q=0.1', 'POT-ML'),
        lty=c(1,1,1,1,1,1,1,1), col=c('grey50','black','red','blue',
        'green','purple', 'cyan3', 'yellow'),text.font=c(2,2,1,1,2,2,1,2))
```

```
#Estimates for k=128
xi_Mom_128<-xi_Mom[128]
xi_Mom_neg_128<-xi_Mom_neg[128]
xi_ghill_128<-xi_ghill[128]
xi_MixMom_128<-xi_MixMom[128]
xi_art_128<-xi_art[128]
xi_PORT_Mom_128<-xi_PORT_Mom[128]
xi_pot_ml_128<-gpd.fit(rpSeg,XX[m-128])$mle[2]
end_Mom_128<-endpoint_Mom_k[128]
end_Mom_neg_128<-endpoint_Mom_neg_k[128]
end_ghill_128<-endpoint_ghill_k[128]
end_MixMom_128<-endpoint_MixMom_k[128]
end_art_128<-endpoint_MOMINV_k[128]
end_PORT_Mom_128<-endpoint_PORT_Mom_k[128]
end_pot_ml_128<-endpoint_pot_ml_k[128]
end_gen_xF_128<-gen_xF[64]

#260<k<270
plot(K+1,endpoint_ghill_k, type='l', xlab='k+1', xlim=c(260,270), ylim=c(540,900),
     ylab=expression(hat(x)[k]*paste('')^F*paste(' (Sec)')),mgp=c(2.5,1,0))
points(K+1,endpoint_Mom_k, type='l', col='red')
points(K+1,endpoint_Mom_neg_k, type='l', col='blue')
points(K+1,endpoint_MixMom_k, type='l', col='green')
points(K+1,endpoint_MOMINV_k, type='l', col='purple')
points(K+1,endpoint_pot_ml_k, type='l', col='yellow')
points(1:NN+1,endpoint_PORT_Mom_k, type='l', col='cyan3')
abline(h=max(XX), lwd=3)
abline(h=max(XX), lwd=3)
abline(v=267, lty=2)
text(267.6,800,expression(k^paste(opt)==266), col='slategrey')
text(269.5,535,expression(X[n:n]==542))
legend(262,900,c('General', 'Generalized Hill','Moment','Negative Moment',
        'Mixed-Moment', 'MOM.inv', 'PORT-Moment q=0.1', 'POT-ML'),
        lty=c(1,1,1,1,1,1,1,1), col=c('grey50','black','red','blue',
        'green','purple', 'cyan3', 'yellow'),text.font=c(2,2,1,1,2,1,1,2))

#Estimates for k=266
xi_Mom_266<-xi_Mom[266]
xi_Mom_neg_266<-xi_Mom_neg[266]
xi_ghill_266<-xi_ghill[266]
xi_MixMom_266<-xi_MixMom[266]
xi_art_266<-xi_art[266]
```

```
xi_PORT_Mom_266<-xi_PORT_Mom[266]
xi_pot_ml_266<-gpd.fit(rpSeg,XX[m-266])$mle[2]
end_Mom_266<-endpoint_Mom_k[266]
end_Mom_neg_266<-endpoint_Mom_neg_k[266]
end_ghill_266<-endpoint_ghill_k[266]
end_MixMom_266<-endpoint_MixMom_k[266]
end_art_266<-endpoint_MOMINV_k[266]
end_PORT_Mom_266<-endpoint_PORT_Mom_k[266]
end_pot_ml_266<-endpoint_pot_ml_k[266]
end_gen_xF_266<-gen_xF[133]

#368<k<380
plot(K+1,endpoint_ghill_k, type='l', xlab='k+1', xlim=c(368,380), ylim=c(540,700),
     ylab=expression(hat(x)[k]*paste('')^F*paste(' (Sec)')),mgp=c(2.5,1,0))
points(K+1,endpoint_Mom_k, type='l', col='red')
points(K+1,endpoint_Mom_neg_k, type='l', col='blue')
points(K+1,endpoint_MixMom_k, type='l', col='green')
points(K+1,endpoint_MOMINV_k, type='l', col='purple')
points(K+1,endpoint_pot_ml_k, type='l', col='yellow')
points(1:NN+1,endpoint_PORT_Mom_k, type='l', col='cyan3')
points(2*K2+1,gen_xF, type='l', col='grey50')
abline(h=max(XX), lwd=3)
abline(v=371, lty=2, lwd=2.5, col='slategrey')
text(371.6,600,expression(k^paste(opt)==370), col='slategrey')
text(379.5,539,expression(X[n:n]==542))
legend(376.5,700,c('General', 'Generalized Hill','Moment','Negative Moment',
        'Mixed-Moment', 'MOM.inv', 'PORT-Moment q=0.1', 'POT-ML'),
        lty=c(1,1,1,1,1,1,1,1), col=c('grey50','black','red','blue',
        'green','purple', 'cyan3', 'yellow'),text.font=c(2,1,1,1,2,1,1,2))

#Estimates for k=370
xi_Mom_370<-xi_Mom[370]
xi_Mom_neg_370<-xi_Mom_neg[370]
xi_ghill_370<-xi_ghill[370]
xi_MixMom_370<-xi_MixMom[370]
xi_art_370<-xi_art[370]
xi_PORT_Mom_370<-xi_PORT_Mom[370]
xi_pot_ml_370<-gpd.fit(rpSeg,XX[m-370])$mle[2]
end_Mom_370<-endpoint_Mom_k[370]
end_Mom_neg_370<-endpoint_Mom_neg_k[370]
end_ghill_370<-endpoint_ghill_k[370]
end_MixMom_370<-endpoint_MixMom_k[370]
```

```
end_art_370<-endpoint_MOMINV_k[370]
end_PORT_Mom_370<-endpoint_PORT_Mom_k[370]
end_pot_ml_370<-endpoint_pot_ml_k[370]
end_gen_xF_370<-gen_xF[185]
```

A.42 Box-Plot Representation of the Yearly Data

```
boxplot(t(dados[,-1]),main='', col='grey90', axes=F, frame.plot=T,
        ylab='Time (Sec)',xlab='Number of Records Set / Year')
axis(1,at=(1:13), labels=anos, las=2)
axis(3,at=(1:13), labels=numb_obs, las=1.5, cex.axis=0.6)
axis(2,labels=T)
```

A.43 Largest 1, 5, 10 and 20 Yearly Observations With Trend Fitting and Plots

```
rfit1_t<-rlarg.fit(dados[,-1],r=1, ydat=dados-2001,mul=c(1))
rfit5_t<-rlarg.fit(dados[,-1],r=5, ydat=dados-2001,mul=c(1))
rfit10_t<-rlarg.fit(dados[,-1],r=10, ydat=dados-2001,mul=c(1))
rfit20_t<-rlarg.fit(dados[,-1],r=20, ydat=dados-2001,mul=c(1))
rlarg.diag(rfit1_t, n=1)
rlarg.diag(rfit5_t, n=5)
rlarg.diag(rfit10_t, n=10)
rlarg.diag(rfit20_t, n=20)

par(mfrow=c(2,2))
plot(anos,dados[,2],type='p',xaxt='n',main='', ylim=c(170,550),
     ylab='Time (Sec)',xlab='Number of Records Set / Year')
axis(1,at=anos, labels=T, las=3)
axis(3,at=anos, labels=numb_obs, las=1.5, cex.axis=0.6)
curve(rfit1_t$mle[1]+rfit1_t$mle[2]*(x-2001),col="grey50",lwd=2,lty=2,add=T)
plot(anos,dados[,2],type='p',xaxt='n',main='', ylim=c(170,550),
     ylab='Time (Sec)',xlab='Number of Records Set / Year')
axis(1,at=anos, labels=T, las=3)
axis(3,at=anos, labels=numb_obs, las=1.5, cex.axis=0.6)
curve(rfit5_t$mle[1]+rfit5_t$mle[2]*(x-2001),col="grey50",lwd=2,lty=2,add=T)
for(i in 3:6){points(anos,dados[,i])}
plot(anos,dados[,2],type='p',xaxt='n',main='', ylim=c(170,550),
     ylab='Time (Sec)',xlab='Number of Records Set / Year')
axis(1,at=anos, labels=T, las=3)
axis(3,at=anos, labels=numb_obs, las=1.5, cex.axis=0.6)
```

```
curve(rfit10_t$mle[1]+rfit10_t$mle[2]*(x-2001),col="grey50",lwd=2,lty=2,add=T)
for(i in 3:11){points(anos,dados[,i])}
plot(anos,dados[,2],type='p',xaxt='n',main='', ylim=c(170,550),
     ylab='Time (Sec)',xlab='Number of Records Set / Year')
axis(1,at=anos, labels=T, las=3)
axis(3,at=anos, labels=numb_obs, las=1.5, cex.axis=0.6)
curve(rfit20_t$mle[1]+rfit20_t$mle[2]*(x-2001),col="grey50",lwd=2,lty=2,add=T)
for(i in 3:21){points(anos,dados[,i])}
par(mfrow=c(1,1))

D1<-2*(-rfit1_t$nllh-(-rfit1$nllh))
D5<-2*(-rfit5_t$nllh-(-rfit5$nllh))
D10<-2*(-rfit10_t$nllh-(-rfit10$nllh))
D20<-2*(-rfit20_t$nllh-(-rfit20$nllh))
Qui2_0.05<-qchisq(0.95,1)

rfit1_t_s<-rlarg.fit(dados[,-1],r=1, ydat=dados-2001,sigl=c(1))
rfit5_t_s<-rlarg.fit(dados[,-1],r=5, ydat=dados-2001,sigl=c(1))
rfit10_t_s<-rlarg.fit(dados[,-1],r=10, ydat=dados-2001,sigl=c(1))
rfit20_t_s<-rlarg.fit(dados[,-1],r=20, ydat=dados-2001,sigl=c(1))
D1_s<-2*(-rfit1_t_s$nllh-(-rfit1$nllh))
D5_s<-2*(-rfit5_t_s$nllh-(-rfit5$nllh))
D10_s<-2*(-rfit10_t_s$nllh-(-rfit10$nllh))
D20_s<-2*(-rfit20_t_s$nllh-(-rfit20$nllh))

rfit1_t_m_s<-rlarg.fit(dados[,-1],r=1, ydat=dados-2001,mul=c(1),sigl=c(1))
rfit5_t_m_s<-rlarg.fit(dados[,-1],r=5, ydat=dados-2001,mul=c(1),sigl=c(1))
rfit10_t_m_s<-rlarg.fit(dados[,-1],r=10, ydat=dados-2001,mul=c(1),sigl=c(1))
rfit20_t_m_s<-rlarg.fit(dados[,-1],r=20, ydat=dados-2001,mul=c(1),sigl=c(1))
D1_m_s<-2*(-rfit1_t_m_s$nllh-(-rfit1_t$nllh))
D5_m_s<-2*(-rfit5_t_m_s$nllh-(-rfit5_t$nllh))
D10_m_s<-2*(-rfit10_t_m_s$nllh-(-rfit10_t$nllh))
D20_m_s<-2*(-rfit20_t_m_s$nllh-(-rfit20_t$nllh))

rfit1_t_x<-rlarg.fit(dados[,-1],r=1, ydat=dados-2001,shl=c(1))
rfit5_t_x<-rlarg.fit(dados[,-1],r=5, ydat=dados-2001,shl=c(1))
rfit10_t_x<-rlarg.fit(dados[,-1],r=10, ydat=dados-2001,shl=c(1))
rfit20_t_x<-rlarg.fit(dados[,-1],r=20, ydat=dados-2001,shl=c(1))
D1_x<-2*(-rfit1_t_x$nllh-(-rfit1$nllh))
D5_x<-2*(-rfit5_t_x$nllh-(-rfit5$nllh))
D10_x<-2*(-rfit10_t_x$nllh-(-rfit10$nllh))
D20_x<-2*(-rfit20_t_x$nllh-(-rfit20$nllh))
```

```
rfit1_t_m_x<-rlarg.fit(dados[,-1],r=1, ydat=dados-2001,mul=c(1),shl=c(1))
rfit5_t_m_x<-rlarg.fit(dados[,-1],r=5, ydat=dados-2001,mul=c(1),shl=c(1))
rfit10_t_m_x<-rlarg.fit(dados[,-1],r=10, ydat=dados-2001,mul=c(1),shl=c(1))
rfit20_t_m_x<-rlarg.fit(dados[,-1],r=20, ydat=dados-2001,mul=c(1),shl=c(1))
D1_m_x<-2*(-rfit1_t_m_x$nllh-(-rfit1_t$nllh))
D5_m_x<-2*(-rfit5_t_m_x$nllh-(-rfit5_t$nllh))
D10_m_x<-2*(-rfit10_t_m_x$nllh-(-rfit10_t$nllh))
D20_m_x<-2*(-rfit20_t_m_x$nllh-(-rfit20_t$nllh))
```

A.44 Largest 1, 5, 10 and 20 Yearly Observations With Trend Estimation

```
prev<-c(14,15,16)
  #r=1
mmo_prob542_r1_t<-NULL ; mmo_endpoint_r1_t<-NULL
mmo_quant0.0001_r1_t<-NULL ; mmo_quant0.01_r1_t<-NULL
for(t in 1:3){
  mmo_prob542_r1_t[t]<-1-pgev(542,rfit1_t$mle[1]+rfit1_t$mle[2]*
        prev[t],rfit1_t$mle[3],rfit1_t$mle[4])
  mmo_quant0.0001_r1_t[t]<-qgev(0.9999,rfit1_t$mle[1]+rfit1_t$mle[2]*
        prev[t],rfit1_t$mle[3],rfit1_t$mle[4])
  mmo_quant0.01_r1_t[t]<-qgev(0.99,rfit1_t$mle[1]+rfit1_t$mle[2]*prev[t],
        rfit1_t$mle[3],rfit1_t$mle[4])}
  #r=5
mmo_prob542_r5_t<-NULL ; mmo_endpoint_r5_t<-NULL
mmo_quant0.0001_r5_t<-NULL ; mmo_quant0.01_r5_t<-NULL
for(t in 1:3){
  mmo_prob542_r5_t[t]<-1-pgev(542,rfit5_t$mle[1]+rfit5_t$mle[2]*
        prev[t],rfit5_t$mle[3],rfit5_t$mle[4])
  mmo_endpoint_r5_t[t]<-(rfit5_t$mle[1]+rfit5_t$mle[2]*prev[t])-
        rfit5_t$mle[3]/rfit5_t$mle[4]
  mmo_quant0.0001_r5_t[t]<-qgev(0.9999,rfit5_t$mle[1]+rfit5_t$mle[2]*
        prev[t],rfit5_t$mle[3],rfit5_t$mle[4])
  mmo_quant0.01_r5_t[t]<-qgev(0.99,rfit5_t$mle[1]+rfit5_t$mle[2]*prev[t],
        rfit5_t$mle[3],rfit5_t$mle[4])}
  #r=10
mmo_prob542_r10_t<-NULL ; mmo_endpoint_r10_t<-NULL
mmo_quant0.0001_r10_t<-NULL ; mmo_quant0.01_r10_t<-NULL
for(t in 1:3){
  mmo_prob542_r10_t[t]<-1-pgev(542,rfit10_t$mle[1]+rfit10_t$mle[2]*
        prev[t],rfit10_t$mle[3],rfit10_t$mle[4])
```

```
  mmo_endpoint_r10_t[t]<-(rfit10_t$mle[1]+rfit10_t$mle[2]*prev[t])-
        rfit10_t$mle[3]/rfit10_t$mle[4]
  mmo_quant0.0001_r10_t[t]<-qgev(0.9999,rfit10_t$mle[1]+rfit10_t$mle[2]*
        prev[t],rfit10_t$mle[3],rfit10_t$mle[4])
  mmo_quant0.01_r10_t[t]<-qgev(0.99,rfit10_t$mle[1]+rfit10_t$mle[2]*prev[t],
        rfit10_t$mle[3],rfit10_t$mle[4])}
  #r=20
mmo_prob542_r20_t<-NULL ; mmo_endpoint_r20_t<-NULL
mmo_quant0.0001_r20_t<-NULL ; mmo_quant0.01_r20_t<-NULL
for(t in 1:3){
  mmo_prob542_r20_t[t]<-1-pgev(542,rfit20_t$mle[1]+rfit20_t$mle[2]*
        prev[t],rfit20_t$mle[3],rfit20_t$mle[4])
  mmo_endpoint_r20_t[t]<-(rfit20_t$mle[1]+rfit20_t$mle[2]*prev[t])-
        rfit20_t$mle[3]/rfit20_t$mle[4]
  mmo_quant0.0001_r20_t[t]<-qgev(0.9999,rfit20_t$mle[1]+rfit20_t$mle[2]*
        prev[t],rfit20_t$mle[3],rfit20_t$mle[4])
  mmo_quant0.01_r20_t[t]<-qgev(0.99,rfit20_t$mle[1]+rfit20_t$mle[2]*prev[t],
        rfit20_t$mle[3],rfit20_t$mle[4])}
```

A.45 NSPOT approach

```
maxleng240<-0
for(i in 1:13){
  if(maxleng240 < length(rpSeg[ANO==2001+i & rpSeg>240]))
    maxleng240<-length(rpSeg[ANO==2001+i & rpSeg>240])}
numb_obs240<-c()
for(i in 1:13){numb_obs240[i]<-length(rpSeg[ANO==2001+i & rpSeg>240])}

dados240<-matrix(rep(NA,times=13*maxleng240),13,maxleng240)
for(i in 1:13){ for(j in 1: numb_obs240[i])
                    dados240[i,j]<-sort(rpSeg[ANO==2001+i & rpSeg>240],
                        decreasing = TRUE)[j]}
dados240<-cbind(anos,dados240)

plot(anos,dados240[,2],ylim=c(min(rpSeg),max(rpSeg)),type='p',xaxt='n',
        main='', ylab='Time (Sec)',xlab='Number of Records Set / Year')
axis(1,at=anos, labels=T, las=3)
axis(3,at=anos, labels=numb_obs240, las=1.5, cex.axis=0.6)
for(i in 2:maxleng240){points(anos,dados240[,i])}
abline(h=240,col='grey50',lty=2, lwd=2)
text(2002.5,235,'4 min', col='grey50')
```

```
th0=240

# -------------- gpd.fit, stationary, sigma=exp(beta0), beta0=log(sigma)
fit_stat_gpd <- gpd.fit(rpSeg,threshold=th0 )
sigma_stat_gpd  <- fit_stat_gpd$mle[1] ;sigma_stat_gpd
beta0_stat_gpd <- log(sigma_stat_gpd);beta0_stat_gpd
gamma_stat_gpd <- fit_stat_gpd$mle[2] ;gamma_stat_gpd
l0_gpd <- -fit_stat_gpd$nllh
xF=th0-sigma_stat_gpd/gamma_stat_gpd

# -------------- gpd.fit, sigma=sigma_t =exp(beta0+beta1*t)
tti=matrix(ncol=1,nrow=length(rpSeg))
tti[,1]=rep(c(1:13),c(numb_obs[1:13]))

aux<-NULL ; obs_ano<-matrix(nrow=1)
for(i in 1:13){
  aux<-sort(dados[i,][is.na(dados[i,])==F][-1])
  obs_ano<-cbind(obs_ano,t(aux))}
obs_ano<-obs_ano[-1]

fit_trend_gpd <- gpd.fit(obs_ano,threshold=th0 ,ydat=tti,sigl=1,siglink=exp)
beta0_gpd <- fit_trend_gpd$mle[1] ;beta0_gpd
beta1_gpd <- fit_trend_gpd$mle[2] ;beta1_gpd
gamma_gpd <- fit_trend_gpd$mle[3] ;gamma_gpd

sigma_t<-function(t){ return(exp(beta0_gpd+t*beta1_gpd))}
nexc <-  fit_trend_gpd$nexc   ;nexc
l1_gpd <- -fit_trend_gpd$nllh ; l1_gpd

# ---- test H=:stationary    vs H1: sigma=sigma_t =exp(beta0+beta1*t)  (trend)
D_gpd = 2*(l1_gpd-l0_gpd) ; D_gpd
1-pchisq(D_gpd,1)

prev<-c(14,15,16)
pot_stat_prob542_240<-(nexc/length(rpSeg))*(1+gamma_stat_gpd*(542-th0)/
        sigma_stat_gpd)^(-1/gamma_stat_gpd)
pot_stat_endpoint_240<-th0-sigma_stat_gpd/gamma_stat_gpd
pot_stat_quant0.0001_240<-th0+sigma_stat_gpd/gamma_stat_gpd*
        ((length(rpSeg)*0.0001/nexc)^-gamma_stat_gpd-1)
pot_stat_quant0.01_240<-th0+sigma_stat_gpd/gamma_stat_gpd*
        ((length(rpSeg)*0.01/nexc)^-gamma_stat_gpd-1)
```

```
pot_prob542_240_t<-NULL ; pot_endpoint_240_t<-NULL
pot_quant0.0001_240_t<-NULL ; pot_quant0.01_240_t<-NULL
for(t in 1:3){
  pot_prob542_240_t[t]<-(nexc/length(rpSeg))*(1+gamma_gpd*(542-th0)/
        sigma_t(prev[t]))^(-1/gamma_gpd)
  pot_endpoint_240_t[t]<-th0-sigma_t(prev[t])/gamma_gpd
  pot_quant0.0001_240_t[t]<-th0+sigma_t(prev[t])/gamma_gpd*
        ((length(rpSeg)*0.0001/nexc)^-gamma_gpd-1)
  pot_quant0.01_240_t[t]<-th0+sigma_t(prev[t])/gamma_gpd*
        ((length(rpSeg)*0.01/nexc)^-gamma_gpd-1)}

detach(female)
```

Bibliography

AIDA (2015). AIDA Rankings. *AIDA International website*. "https://www.aidainternational.org/ranking".

AIDA (2016). AIDA competitive freediving. *AIDA International website*. "https://www.aidainternational.org/competitive".

Araújo Santos, P., Fraga Alves, M. I., and Gomes, M. I. (2006). Peaks over random threshold methodology for tail index and high quantile estimation. *REVSTAT – Statistical Journal*, 4(3):227–247.

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (2008). *A First Course in Order Statistics*. Society for Industrial and Applied Mathematics, United States of America.

Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5):792–804.

Beguería, S., Angulo-Martínez, M., Vicente-Serrano, S. M., López-Moreno, J. I., and El-Kenawyb, A. (2011). Assessing trends in extreme precipitation events intensity and magnitude using non-stationary peaks-over-threshold analysis: a case study in northeast Spain from 1930 to 2006. *International Journal of Climatology*, 31:2102–2114.

Beirlant, J., Dierckx, G., and Guillou, A. (2005). Estimation of the extreme-value index and generalized quantile plots. *Bernoulli*, 11(6):949–970.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. John Wiley & Sons, England.

Beirlant, J., Vynckier, P., and Teugels, J. (1996). Excess functions and estimation of the extreme-value index. *Bernoulli*, 2(4):293–318.

Bermudez, P. Z. and Kotz, S. (2010a). Parameter estimation of the generalized Pareto distribution – Part I. *Journal of Statistical Planning and Inference*, 140(6):1353–1373.

Bermudez, P. Z. and Kotz, S. (2010b). Parameter estimation of the generalized Pareto distribution – Part II. *Journal of Statistical Planning and Inference*, 140(6):1374–1388.

Bermudez, P. Z. and Turkman, M. A. A. (2003). Bayesian approach to parameter estimation of the generalized Pareto distribution. *Test*, 12(1):259–277.

Birnbaum, Z. W. (1953). Distribution-free tests of fit for continuous distribution functions. *The Annals of Mathematical Statistics*, 24(1):1–8.

Cabaña, A. and Quiroz, A. J. (2005). Using the empirical moment generating function in testing for the Weibull and the type I extreme value distributions. *Socíedad de Estadística e Investigación Operativa*, 14(2):417–431.

Caeiro, F. and Gomes, M. I. (2010). An asymptotically unbiased moment estimator of a negative extreme value index. *Discussiones Mathematica: Probability and Statistics*, 30:5–19.

Castillo, E., Hadi, A. S., Balakrishnan, N., and Sarabia, J. M. (2004). *Extreme Value and Related Models with Applications in Engineering and Science.* John Wiley & Sons, New York.

Chandra, M., Singpurwalla, N. D., and Stephens, M. A. (1981). Kolmogorov statistics for tests of fit for the extreme value and weibull distributions. *Journal of the American Statistical Association*, 76(375):729–731.

Chavez-Demoulin, V. and Roehrl, A. (2004). Extreme Value Theory can save your neck. *ETHZ.*

Choulakian, V. and Stephens, M. A. (2001). Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*, 43(4):478–484.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer, London.

Davison, A. (1984). *Statistical Extremes and Applications*, chapter Modeling excesses over high threshold with an application, pages 461–482. D. Reidel.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society, Series B*, 52(3):393–442.

de Haan, L. (1970). *On Regular Variation and its Applications to the Weak Convergence of Sample Extremes.* Mathematical Centre Tract, 32$^{\text{nd}}$ edition.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory – An Introduction.* Springer, New York.

de Haan, L., Klein Tank, A., and Neves, C. (2015). On tail trend detection: modeling relative risk. *Extremes*, 18:141–178.

de Haan, L. and Stadtmüller, U. (1996). Generalized Regular Variation of Second Order. *Journal of the Australian Mathematical Society*, (A61):381–395.

Dekkers, A. L. M. and de Haan, L. (1989). On the estimation of the extreme-value index and large quantile estimation. *Annals of Statistics*, 17:1795–1832.

Dekkers, A. L. M., Einmahl, J. H. J., and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *Annals of Statistics*, 17:1833–1855.

Drees, H., Ferreira, A., and de Haan, L. (2004). On maximum likelihood estimation of the extreme value index. *Annals of Applied Probability*, 14:1179–1201.

Einmahl, J. H. J., de Haan, L., and Sinha, A. K. (1997). Estimating the spectral measure of an extreme value distribution. *Stochastic Processes and their Applications*, 70:143–171.

Einmahl, J. H. J., de Haan, L., and Zhou, C. (2016). Statistics of heteroscedastic extremes. *Journal of the Royal Statistical Society – Series B*, 78(1):31–51.

Engineering Sport (2016). How long can we hold our breath for? *Engineering Sport website*. "https://engineeringsport.co.uk/2012/06/06/how-long-can-we-hold-our-breath-for/".

Falk, M. (1995). Some best estimators for distributions with finite endpoint. *Statistics*, 27:115–125.

Fawcett, L. (2012). Topics in Statistics: Environmental Extremes. Lecture notes, Newcastle University – School of Mathematics & Statistics.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, volume 2. John Wiley & Sons, New York.

Ferreira, A., de Haan, L., and Peng, L. (2003). On optimising the estimation of high quantiles of a probability distribution. *Statistics: A Journal of Theoretical and Applied Statistics*, 37(5):401–434.

Fischer, H. (2011). *A History of the Central Limit Theorem – From Classical to Modern Probability Theory*. Springer, New York.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Cambridge Philos. Soc.*, 24:180–190.

Fraga Alves, M. I. (1992, 1995). Estimation of the tail parameter in the domain of attraction of an extremal distribution. *Extreme Value Theory and Applications* (Villeneuve d'Ascq, 1992) and *Journal of Statistical Planning and Inference*, 45(1-2):143–173.

Fraga Alves, M. I. (1999). Asymptotic distribution of Gumbel statistic in a semi-parametric approach. *Portugaliae Mathematica*, 56(3):281–298.

Fraga Alves, M. I. (2001). A location invariant Hill-type estimator. *Extremes*, 4:199–217.

Fraga Alves, M. I. (2015). *Mathematics of Energy and Climate Change*, volume 2, chapter Max-stability at work (or not): estimating return levels for daily rainfall data, pages 1–14. Springer International Publishing.

Fraga Alves, M. I., de Haan, L., and Neves, C. (2009a). A test procedure for detecting super-heavy tails. *Journal of Statistical Planning and Inference*, 139:213–227.

Fraga Alves, M. I., Gomes, M. I., de Haan, L., and Neves, C. (2009b). Mixed moment estimator and location invariant alternatives. *Extremes*, 12:149–185.

Fraga Alves, M. I. and Neves, C. (2014). Estimation of the finite right endpoint in the Gumbel domain. *Statistica Sinica*, 24:1811–1835.

Fraga Alves, M. I. and Neves, C. (2015). *Handbook of Extreme Value Thory and Applications in Finance*, chapter Extreme Value Theory: An Introductory Overview. John Wiley & Sons, 1st edition.

Fraga Alves, M. I., Neves, C., and Cormann, U. (2011). *Laws of Small Numbers: Extremes and Rare Events*, chapter Heavy and Super-Heavy Tail Analysis, pages 75–101. Birkhäuser Basel, 3rd edition.

Fraga Alves, M. I., Neves, C., and Rosário, P. (2016). A general estimator for the right endpoint with an application to supercentenarian women's records. Extremes:1–39.

Fraga Alves, M. I. and Rosário, P. (2015). *Theory and Practice of Risk Assessment*, chapter Parametric and Semi-parametric Approaches to Extreme Rainfall Modelling, pages 279–291. Springer International Publishing.

Fréchet, M. (1927). Sur la loi de probabilite de l'écart maximum. *Annales de la Société Polonaise de Mathématique*, 6:93–116.

Freedive UK (2016). How to hold your breath for 5 minutes in 1 month – Freediving training. *Freedive UK website*. "`http://freediveuk.com/how-to-hold-your-breath-for-5-minutes-in-1month-freediving-training`".

Gnedenko, B. (1943). Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire. *Annals of Mathematics*, 44(3):423–453.

Gomes, M. I. (1981). *Statistical distributions in Scientific Work*, volume 6, chapter An i-dimensional limiting distribution function of largest values and its relevance to the statistical theory of extremes, pages 389–410. D. Reidel.

Gomes, M. I. and Fraga Alves, M. I. (1996). Statistical choice of extreme value domains of attraction - a comparative analysis. *Communications in Statistics – Part A: Theory and Methods*, 25(4):789–811.

Gomes, M. I., Fraga Alves, M. I., and Neves, C. (2013a). *Análise de Valores Extremos – Uma Introdução*. Sociedade Portuguesa de Estatística.

Gomes, M. I., Fraga Alves, M. I., and Santos, P. A. (2008). PORT Hill and moment estimators for heavy-tailed models. *Communications in Statistics – Simulation and Computation*, 37(7):1281–1306.

Gomes, M. I., Henriques-Rodrigues, L., and Caeiro, F. (2013b). *Advances in Theoretical and Applied Statistics*, chapter Refined estimation of a light tail: an application to environmental data, pages 143–153. Springer Berlin Heidelberg.

Gomes, M. I. and Oliveira, O. (2001). The bootstrap methodology in statistics of extremes – choice of the optimal sample fraction. *Extremes*, 4(4):331–358.

Gomes, M. I. and van Monfort, M. A. J. (1986). *Procedures of the III International Conference on Statistical Climatology*, chapter Exponentiality versus Generalized Pareto, quick tests, pages 185–195.

Gong, S. (2012). Estimation of hot and cold spells with extreme value theory. Project Report 19, Uppsala University.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water Resourses Research*, 15(5):1049–1054.

Greenwood, M. (1946). The statistical study of infectious diseases. *Journal of the Royal Statistical Society*, 109(2):85–110.

Gumbel, E. J. (1985). *Statistics of Extremes*. Columbia University Press, New York.

Hasofer, A. M. and Wang, Z. (1992). A test for extreme value domain of attraction. *Journal of the American Statistical Association*, 87(417):171–177.

Henriques-Rodrigues, L., Gomes, M. I., and Pestana, D. (2011). Statistics of extremes in athletics. *REVSTAT – Statistical Journal*, 9(2):127–153.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3(5):1163–1174.

Hosking, J. R. M. (1984). Testing whether the shape parameter is zero in the generalized extreme value distribution. *Biometrika*, 71(2):367–374.

Hosking, J. R. M. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29:339–349.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27:251–261.

Hüsler, J. and Peng, L. (2008). Review of testing issues in extremes: in honor of Professor Laurens de Haan. *Extremes*, 11:99–111.

Immersion Freediving (2016). PFI Freediver Course. *Immersion Freediving website*. "http://immersionfreediving.com/classes-2/2-andahalf-day-class/".

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of metereological elements. *Quarterly Journal of the Royal Metereology Society*, 81:158–171.

Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability weighted moments compared with some traditional techniques in estimating gumbel parameters and quantiles. *Water Resourses Research*, 15:1055–1064.

Lang, M., Ouarda, T. B. M. J., and Bobée, B. (1999). Towards operational guidelines for over-threshold modeling. *Journal of Hydrology*, 225:103–117.

Lilliefors, H. W. (1969). On the Kolmogorov-Smirnov test for the Exponential distribution with mean unknown. *Journal of the American Statistical Association*, 64(325):387–389.

Marohn, F. (1998). Testing the Gumbel hypothesis via the Pot-method. *Extremes*, 1(2):191–213.

Marohn, F. (2000). Testing extreme value models. *Extremes*, 3(4):363–384.

Méndez, F. J., Menéndez, M., Luceño, A., and Losada, I. J. (2006). Estimation of the long-term variability of extreme significant wave height using a time-dependent Peak Over Threshold (POT) model. *Journal OF Geophysical Research*, 111.

Neves, C. and Fraga Alves, M. I. (2007). Semi-parametric approach to the Hasofer-Wang and Greenwood statistics in extremes. *Test*, 16:297–313.

Neves, C. and Fraga Alves, M. I. (2008). Testing extreme value conditions – An overview and recent approaches. *REVSTAT – Statistical Journal*, 6(1):83–100.

Neves, C. and Pereira, A. (2010). Detecting finiteness in the right endpoint of light-tailed distributions. *Statistics and Probability Letters*, 80:437–444.

Neves, C., Picek, J., and Fraga Alves, M. I. (2006). The contribution of the maximum to the sum of excesses for testing max-domains of attraction. *Journal of Statistical Planning and Inference*, 136(4):1281–1301.

Nogaj, M., Parey, S., and Dacunha-Castelle, D. (2007). Non-stationary extreme models and a climatic application. *Nonlinear Processes in Geophysics*, 14:305–316.

Nogaj, M., Yiou, P., Parey, S., Malek, F., and Naveau, P. (2006). Amplitude and frequency of temperature extremes over the North Atlantic region. *Geophysical Research Letters*, 33(10).

Peng, L. (1998). Asymptotically unbiased estimators for extreme value index. *Statistics & Probability Letters*, 38:107–115.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131.

Razali, N. M. and Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33.

Reiss, R. D. and Thomas, M. (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhäuser, Germany, 3rd edition.

Renard, B., Lang, M., and Bois, P. (2006). *Stochastic Environmental Research and Risk Assessment*, chapter Statistical analysis of extreme events in a nonstationary context via a Bayesian framework. Case study with peak-over-threshold data, pages 97–112. Springer Verlag, Germany.

Rosário, P. (2013). Análise de valores extremos para níveis pluviométricos em barcelos. Master's thesis, Faculdade de Ciências da Universidade de Lisboa, Lisbon.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90.

Smith, R. L. (1986). Extreme value theory based on the r largest annual events. *Journal of Hydrology*, 86:27–43.

Smith, R. L. (1987). Estimating tails of probability distributions. *Annals of Statistics*, 15:1174–120.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, 4(4):367–377.

Stephens, M. A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *Annals of Statistics*, 4:357–369.

Stephens, M. A. (1977). Goodness-of-fit for the extreme value distribution. *Biometrika*, 64(3):583–588.

Stephens, M. A. (1986). *Goodness-of-Fit Techniques*, chapter Tests for the Exponential Distribution, pages 421–459. Marcel Dekker, Inc.

Stephenson, A. G. and Twan, J. A. (2013). Determining the best track performances of all time using a conceptual population model for athletic records. *Journal of Quantitative Analysis in Sports*, 9(1):67–76.

Themido Pereira, T. (1993). *Extreme Value Theory and Applications III, Proc. Gaithersburg Conference*, chapter Second order behaviour of domains of attraction and the bias of generalized Pickands' estimator, pages 165–177. National Institute of Standards and Technology special publication.

Tiago de Oliveira, J. (1981). *Statistical distributions in Scientific Work*, volume 6, chapter Statistical choice of univariate extreme models, pages 367–387. D. Reidel.

Tiago de Oliveira, J. and Gomes, M. I. (1984). *Statistical Extremes and Applications*, chapter Two test statistics for choice of univariate extreme models, pages 651–668. D. Reidel.

Vanem, E. (2015). Non-stationary extreme value models to account for trends and shiftsin the extreme wave climate due to climate change. *Applied Ocean Research*, 52:201–211.

Vicente, S. (2012). Extreme value theory: an application to sports. Master's thesis, Faculdade de Ciências da Universidade de Lisboa, Lisbon.

von Mises, R. (1936). La distribution de la plus grande de n valeurs. *Revue Mathématique de l'Union. Interbalcanique*, 1:141–160.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364):812–815.

Yun, S. (2002). On a generalized Pickands estimator of the extreme value index. *Journal of Statistical Planning and Inference*, 102:389–409.

Zhou, C. (2009). Existence and consistency of the maximum likelihood estimator for the extreme value index. *Journal of Multivariate Analysis*, 100(4):794–815.

Zhou, C. (2010). The extent of the maximum likelihood estimator for the extreme value index. *Journal of Multivariate Analysis*, 101(4):971–983.